



US009317453B2

(12) **United States Patent**
Armstrong et al.

(10) **Patent No.:** **US 9,317,453 B2**
(45) **Date of Patent:** **Apr. 19, 2016**

(54) **CLIENT PARTITION SCHEDULING AND
PRIORITIZATION OF SERVICE PARTITION
WORK**

(56) **References Cited**

U.S. PATENT DOCUMENTS

(75) Inventors: **William J. Armstrong**, Rochester, MN
(US); **Orran Y. Krieger**, Newton, MA
(US); **Michal Ostrowski**, Austin, TX
(US); **Randal C. Swanberg**, Round
Rock, TX (US)

5,432,929 A 7/1995 Escola et al.
5,555,385 A 9/1996 Osisek

(Continued)

FOREIGN PATENT DOCUMENTS

(73) Assignee: **INTERNATIONAL BUSINESS
MACHINES CORPORATION**,
Armonk, NY (US)

EP 1548589 A2 6/2005
EP 1691287 A1 8/2006
JP 2008041093 A 2/2008

OTHER PUBLICATIONS

(*) Notice: Subject to any disclaimer, the term of this
patent is extended or adjusted under 35
U.S.C. 154(b) by 926 days.

Armstrong et al., Notice of Allowance for U.S. Appl. No. 13/441,139,
filed Apr. 6, 2013 (U.S. Patent Publication No. 2012/0210044 A1),
dated Mar. 25, 2013.

(Continued)

(21) Appl. No.: **13/349,071**

(22) Filed: **Jan. 12, 2012**

(65) **Prior Publication Data**

US 2012/0117353 A1 May 10, 2012

Related U.S. Application Data

(63) Continuation of application No. 12/110,953, filed on
Apr. 28, 2008, now Pat. No. 8,176,487.

(60) Provisional application No. 60/953,512, filed on Aug.
2, 2007.

(51) **Int. Cl.**
G06F 9/455 (2006.01)
G06F 12/14 (2006.01)

(52) **U.S. Cl.**
CPC **G06F 12/1475** (2013.01); **G06F 9/45558**
(2013.01); **G06F 12/1491** (2013.01); **G06F**
2009/45579 (2013.01); **G06F 2212/1056**
(2013.01)

(58) **Field of Classification Search**

None

See application file for complete search history.

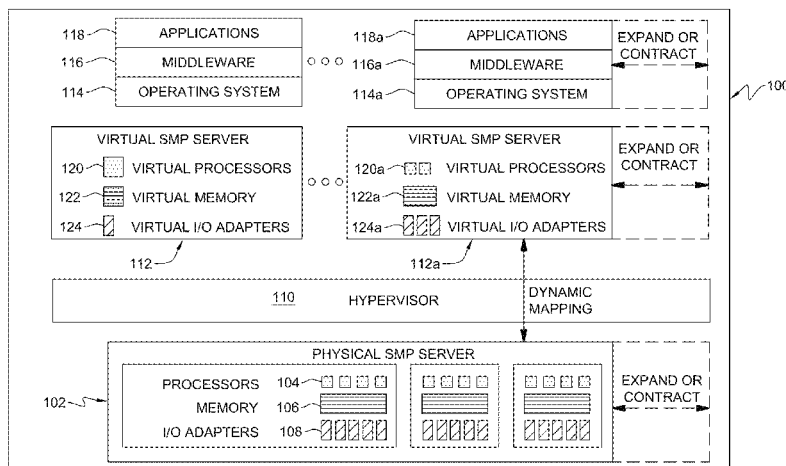
Primary Examiner — Jacob A Petranek

(74) *Attorney, Agent, or Firm* — Parashos Kalaitzis, Esq.;
Kevin P. Radigan, Esq.; Heslin Rothenberg Farley & Mesiti
P.C.

(57) **ABSTRACT**

A method in a data processing system is provided for processing a service request of a client partition. The method includes: obtaining by a service partition of the data processing system the service request from the client partition, wherein both the client and service partitions execute above a hypervisor of the data processing system; and processing the service request by the service partition utilizing a processor quantum assigned to the client partition and donated by the client partition to the service partition. The client partition controls scheduling of the service partition by queuing the service request at the client partition until the client partition decides to proceed with execution of the service request by the service partition. In one implementation, the service partition is a partition adjunct of the data processing system, which utilizes donated virtual address space of the client partition.

20 Claims, 24 Drawing Sheets



(56)

References Cited

U.S. PATENT DOCUMENTS

5,729,723 A 3/1998 Wada et al.
 6,256,704 B1 7/2001 Hlava et al.
 6,473,655 B1 10/2002 Gould et al.
 6,629,162 B1 9/2003 Arndt et al.
 6,892,383 B1 5/2005 Arndt
 6,901,497 B2 5/2005 Tashiro et al.
 6,970,990 B2 11/2005 Rogers et al.
 7,073,002 B2 7/2006 Armstrong et al.
 7,076,634 B2 7/2006 Lambeth et al.
 7,296,133 B2 11/2007 Swanberg
 7,617,377 B2 11/2009 Moertl et al.
 7,802,081 B2 9/2010 Schultz et al.
 8,010,763 B2 8/2011 Armstrong et al.
 2002/0016892 A1 2/2002 Zalewski et al.
 2003/0037089 A1 2/2003 Cota-Robles et al.
 2004/0194096 A1 9/2004 Armstrong et al.
 2004/0215917 A1 10/2004 Lambeth et al.
 2005/0027941 A1 2/2005 Wang et al.
 2005/0071841 A1 3/2005 Hoflehner et al.
 2005/0125802 A1 6/2005 Wang et al.
 2006/0020943 A1 1/2006 Boutcher et al.
 2006/0047959 A1 3/2006 Morais
 2006/0064523 A1 3/2006 Moriki et al.
 2006/0090058 A1 4/2006 Chng et al.
 2006/0095690 A1 5/2006 Craddock et al.
 2006/0143417 A1 6/2006 Poisner et al.
 2006/0195663 A1 8/2006 Arndt et al.
 2006/0209863 A1 9/2006 Arndt et al.
 2006/0230219 A1 10/2006 Njoku et al.
 2006/0248527 A1 11/2006 Jaekel et al.
 2006/0250945 A1 11/2006 Fernandes et al.
 2006/0259732 A1 11/2006 Traut et al.
 2006/0282624 A1 12/2006 Yokota
 2006/0294288 A1 12/2006 Seth et al.
 2007/0050764 A1 3/2007 Traut
 2007/0169005 A1 7/2007 Drepper
 2007/0169024 A1 7/2007 Drepper
 2007/0180210 A1 8/2007 Thibadeau
 2007/0220246 A1 9/2007 Powell et al.
 2007/0271565 A1 11/2007 Tirumalai et al.
 2008/0016337 A1 1/2008 Morgan et al.
 2008/0091855 A1 4/2008 Moertl et al.
 2008/0126690 A1 5/2008 Rajan et al.
 2008/0307425 A1 12/2008 Tripathi

2009/0037906 A1 2/2009 Armstrong et al.
 2009/0037907 A1 2/2009 Armstrong et al.
 2009/0037908 A1 2/2009 Armstrong et al.
 2009/0037941 A1 2/2009 Armstrong et al.

OTHER PUBLICATIONS

Armstrong et al., Notice of Allowance for U.S. Appl. No. 12/110,953, filed Apr. 28, 2008 (U.S. Patent Publication No. 2009/0037907 A1), dated Dec. 29, 2011.
 Armstrong et al., Notice of Allowance for U.S. Appl. No. 12/110,923, filed Apr. 28, 2008 (U.S. Patent Publication No. 2009/0037906 A1), dated Mar. 2, 2012.
 Armstrong et al., Notice of Allowance for U.S. Appl. No. 12/110,982, filed Apr. 28, 2008 (U.S. Patent Publication No. 2009/0037908 A1), dated Mar. 2, 2012.
 Nakano, "Virtualization, Looking Closely Into the Reality of Virtualization," UNIX Magazine, Japan, ASCII Corporation, Jan. 1, 2007, vol. 22, No. 1, pp. 46-53.
 Miyazaki, "Secure OS/Trusted OS-Looking Into the Secure Functions of OS," UNIX Magazine, Japan, ASCII Corporation, Apr. 1, 2007, vol. 22, No. 2, pp. 36-43.
 Translations of portions of Nov. 15, 2011 Office Action that apply, discuss, or otherwise mention the attached references.
 "Single Root I/O Virtualization and Sharing Specification—Revision 1.0", PCI-SIG (Sep. 11, 2007).
 "Xen: Enterprise Grade Open Source Virtualization, Inside Xen 3.0—A XenSource White Paper", XenSource, Inc., (pp. 1-4) (Jun. 1, 2006).
 Office Action for U.S. Appl. No. 12/110,923 (US Application Publication No. 2009-0037906 A1) dated Oct. 6, 2011.
 Office Action for U.S. Appl. No. 12/110,953 (US Application Publication No. 2009-0037907 A1) dated Aug. 11, 2011.
 Final Office Action for U.S. Appl. No. 12/110,953 (US Application Publication No. 2009-0037907 A1) dated Nov. 17, 2011.
 Office Action for U.S. Appl. No. 12/110,982 (US Application Publication No. 2009-0037908 A1) dated Sep. 29, 2011.
 Office Action for U.S. Appl. No. 12/111,020 (US Application Publication No. 2009-0037941 A1) dated Jul. 20, 2011.
 Office Action for U.S. Appl. No. 12/111,041 (US Application Publication No. 2009-0037682 A1) dated Nov. 19, 2010.
 Final Office Action for U.S. Appl. No. 12/111,020 (US Application Publication No. 2009-0037941 A1) dated Jan. 4, 2012.

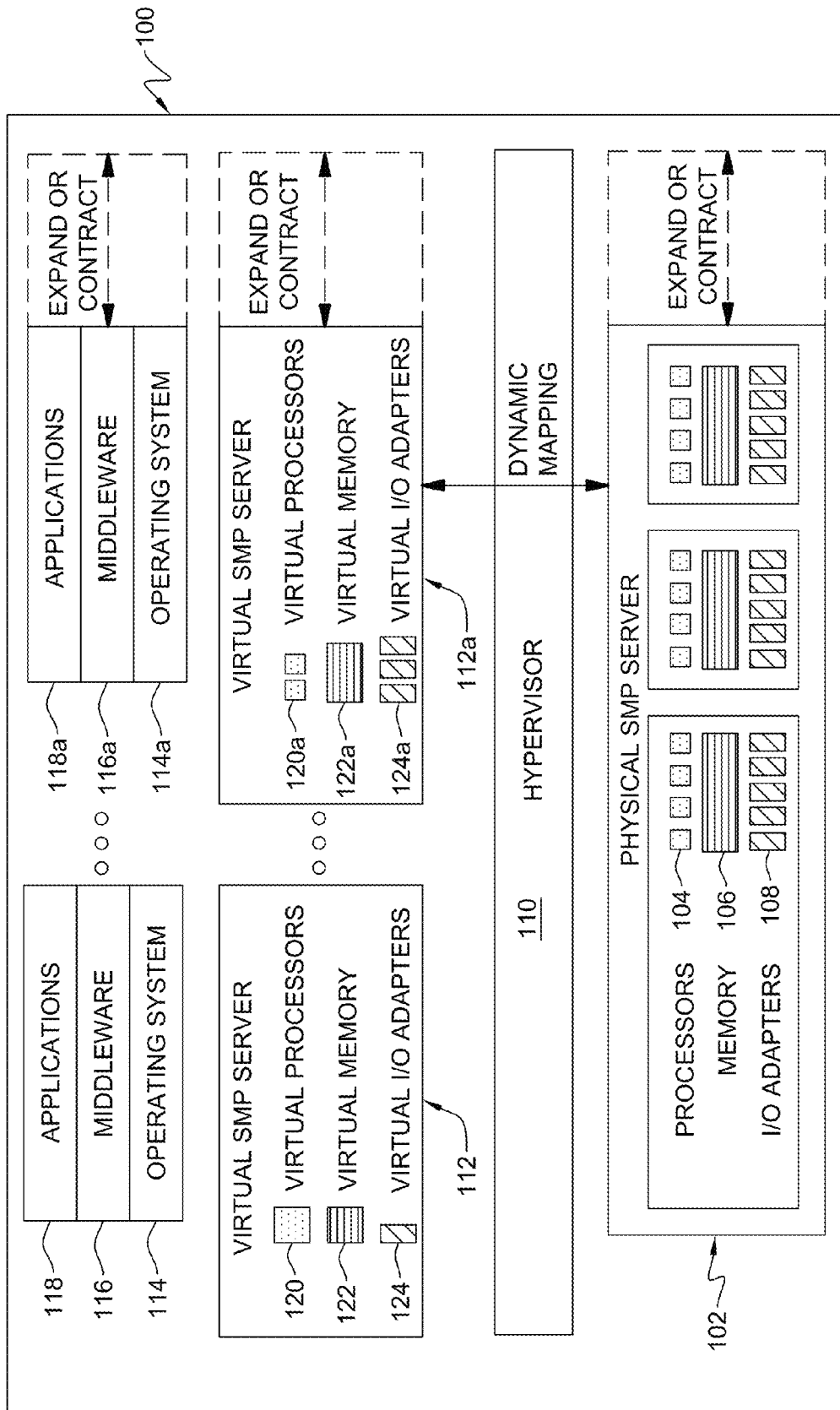


FIG. 1

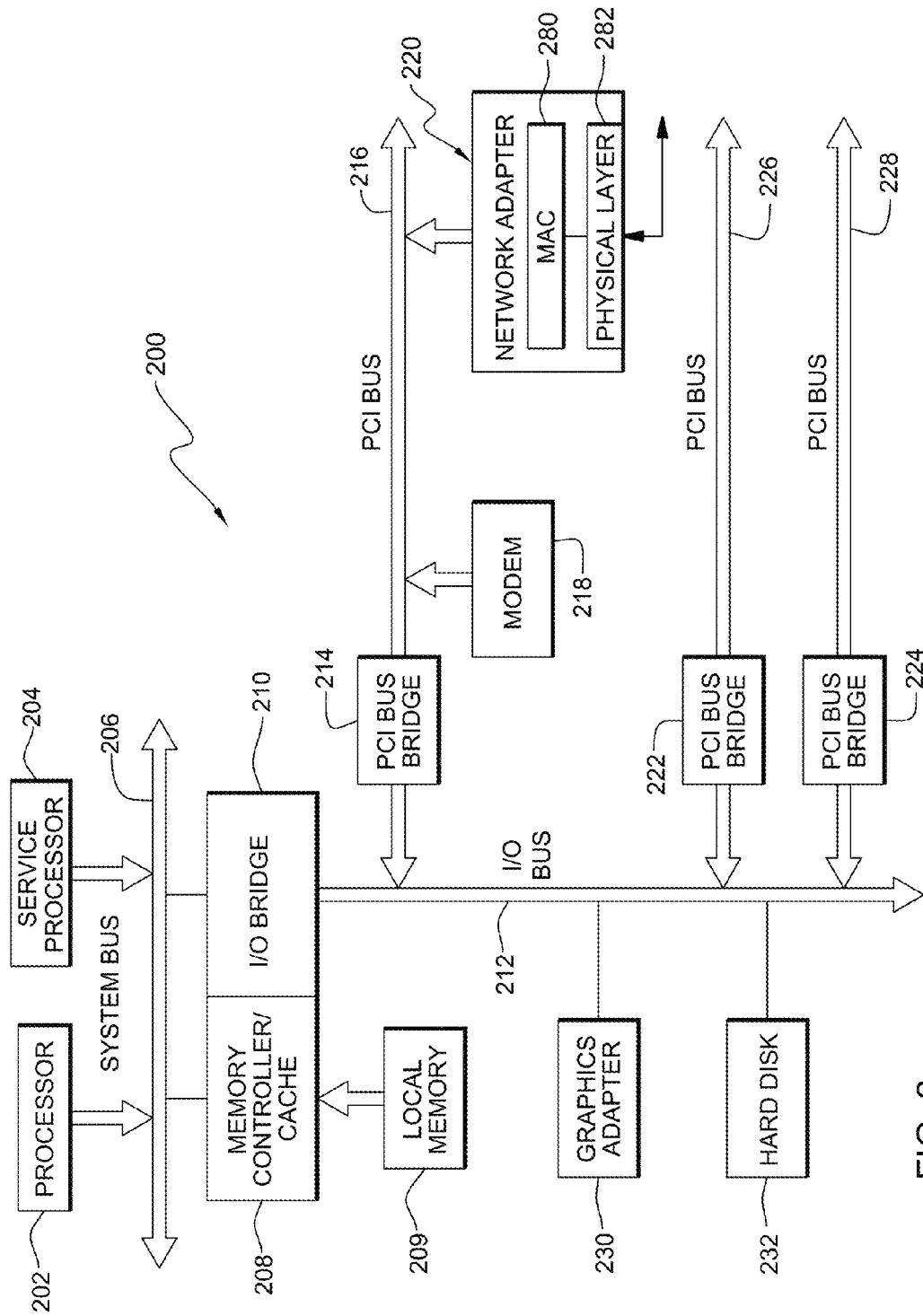


FIG. 2

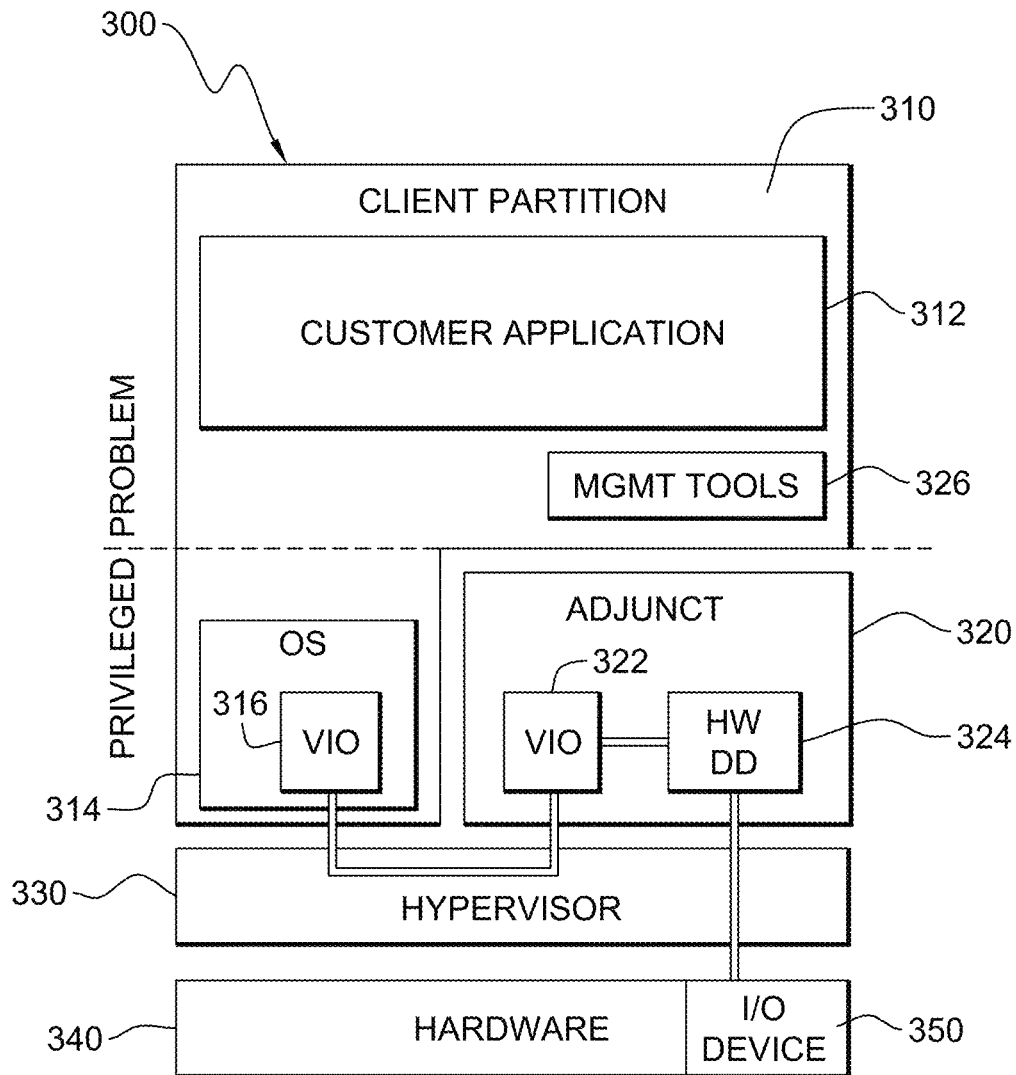


FIG. 3

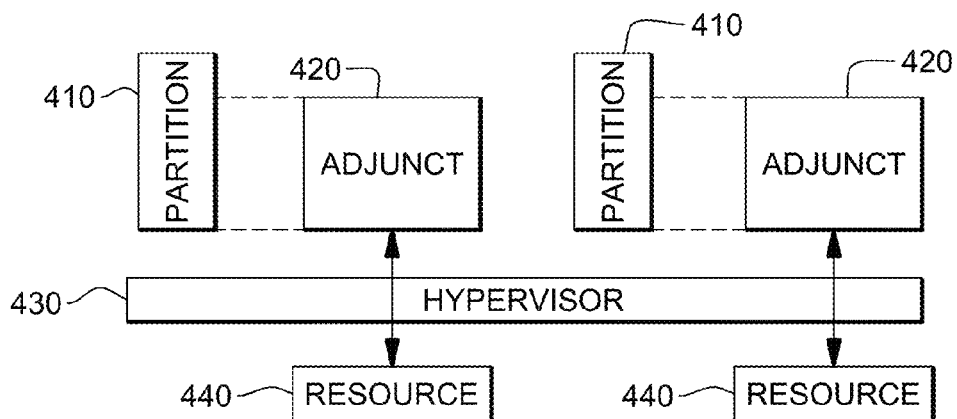


FIG. 4A

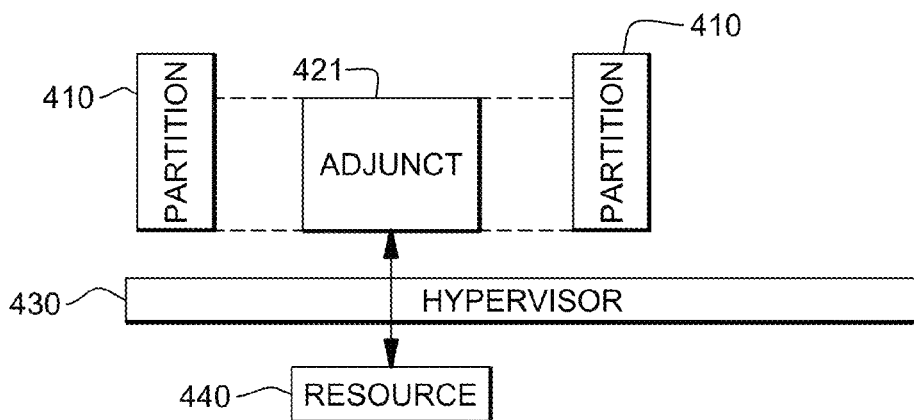


FIG. 4B

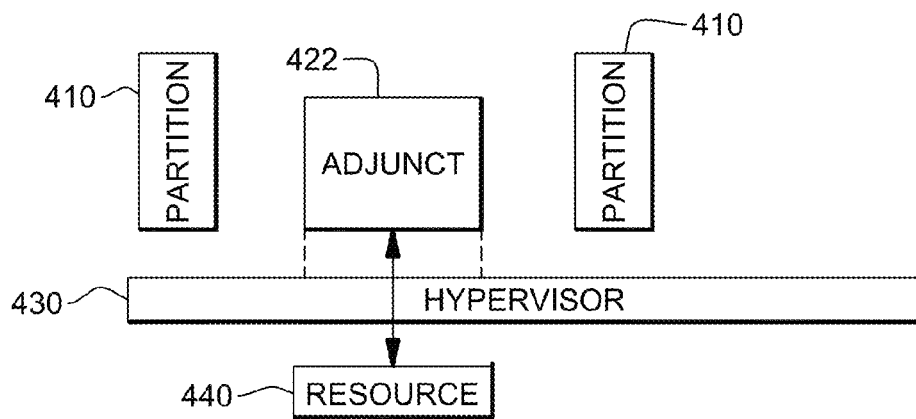


FIG. 4C

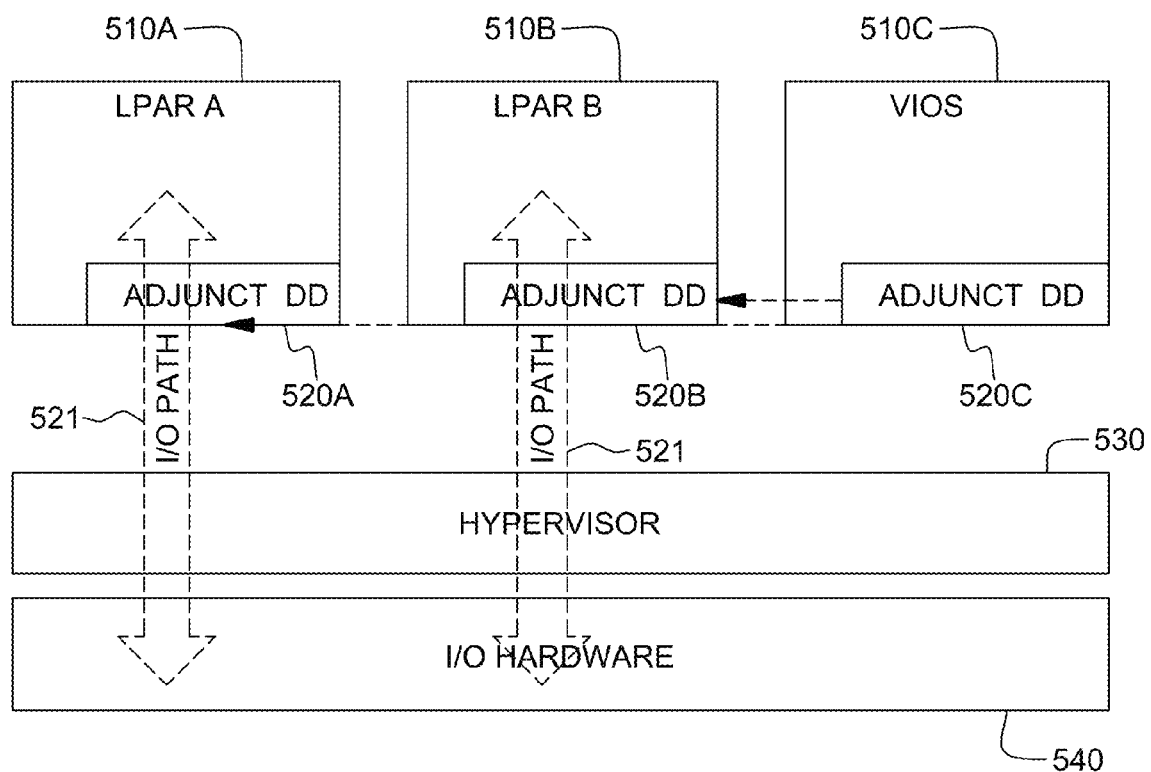


FIG. 5

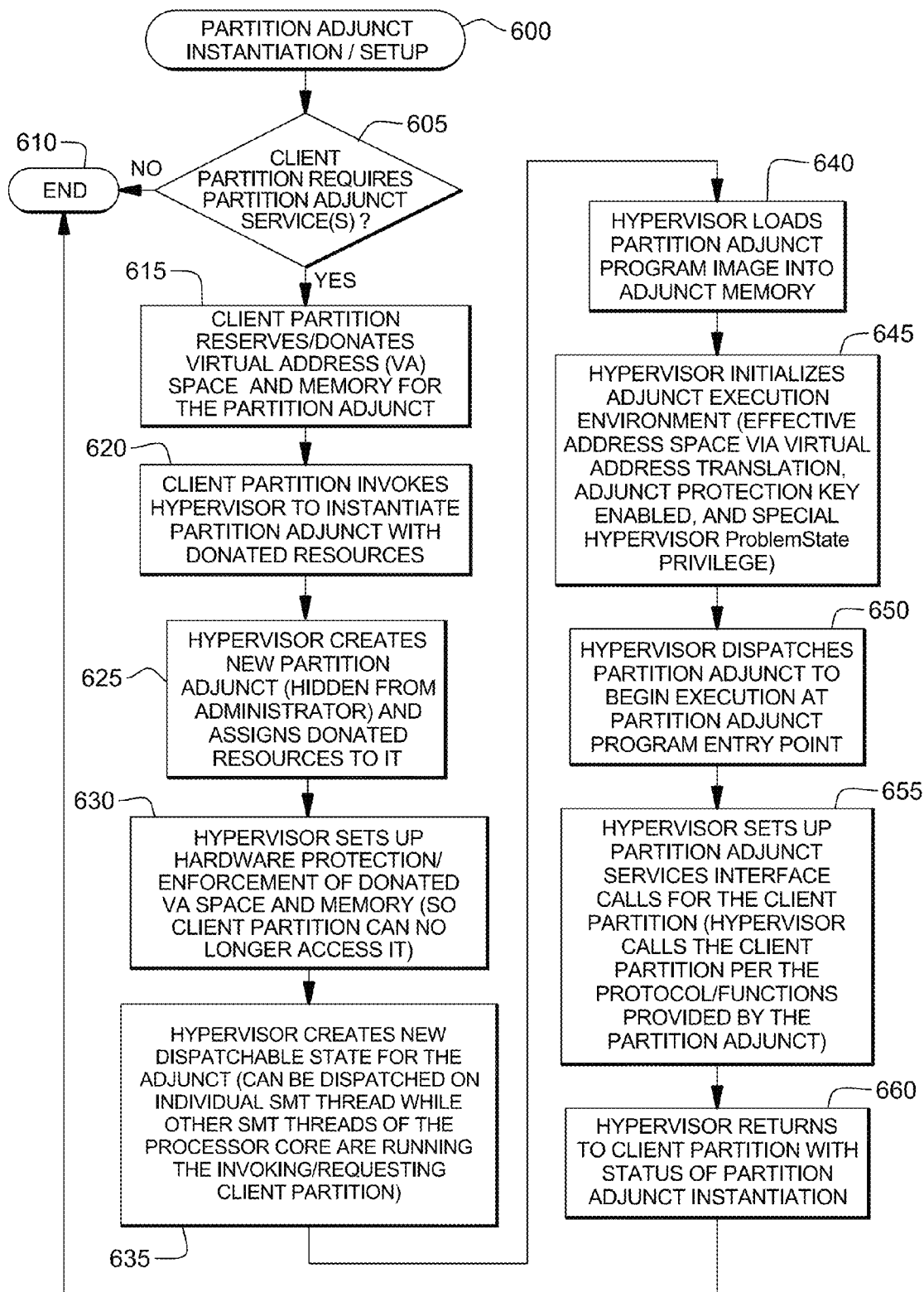
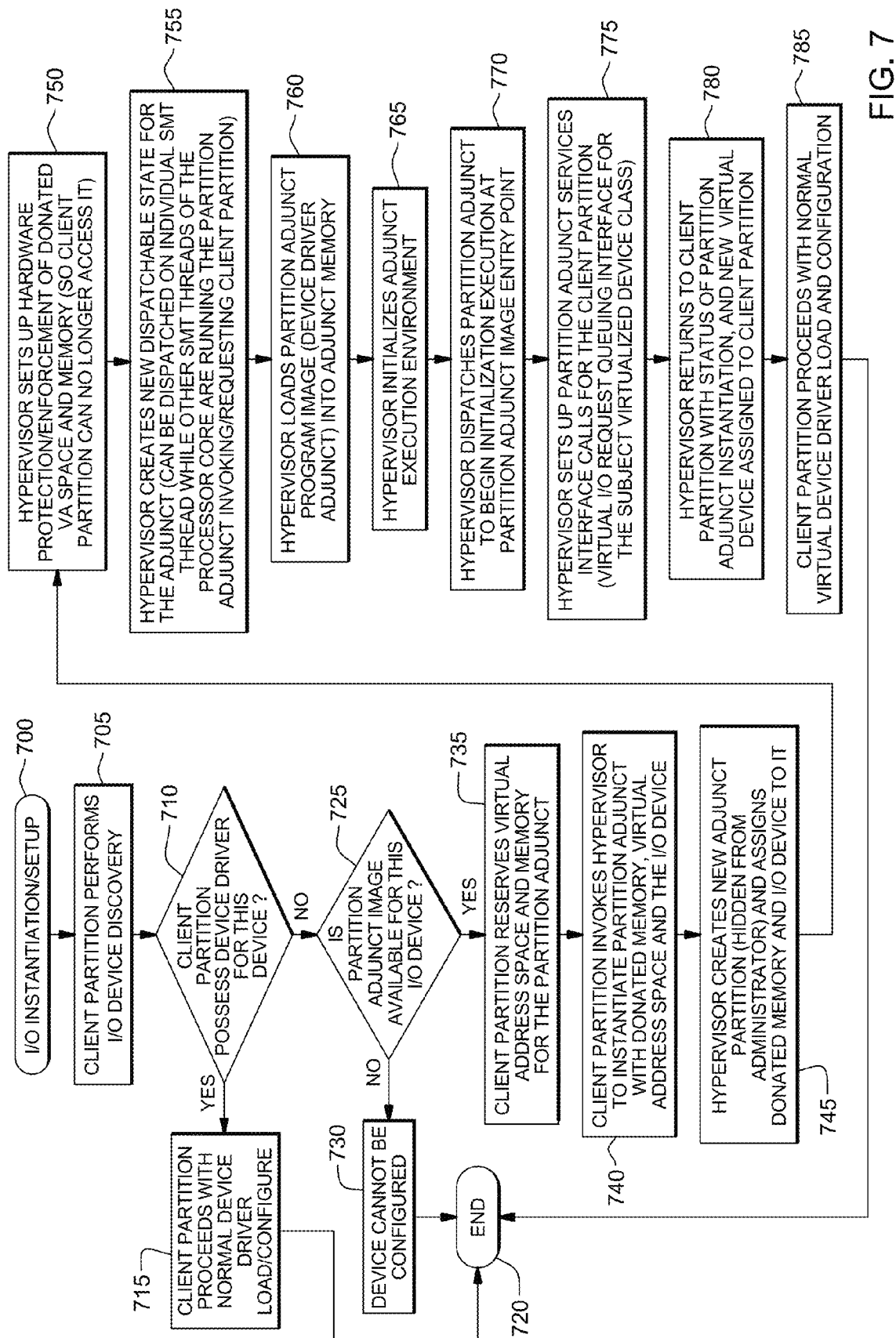


FIG. 6



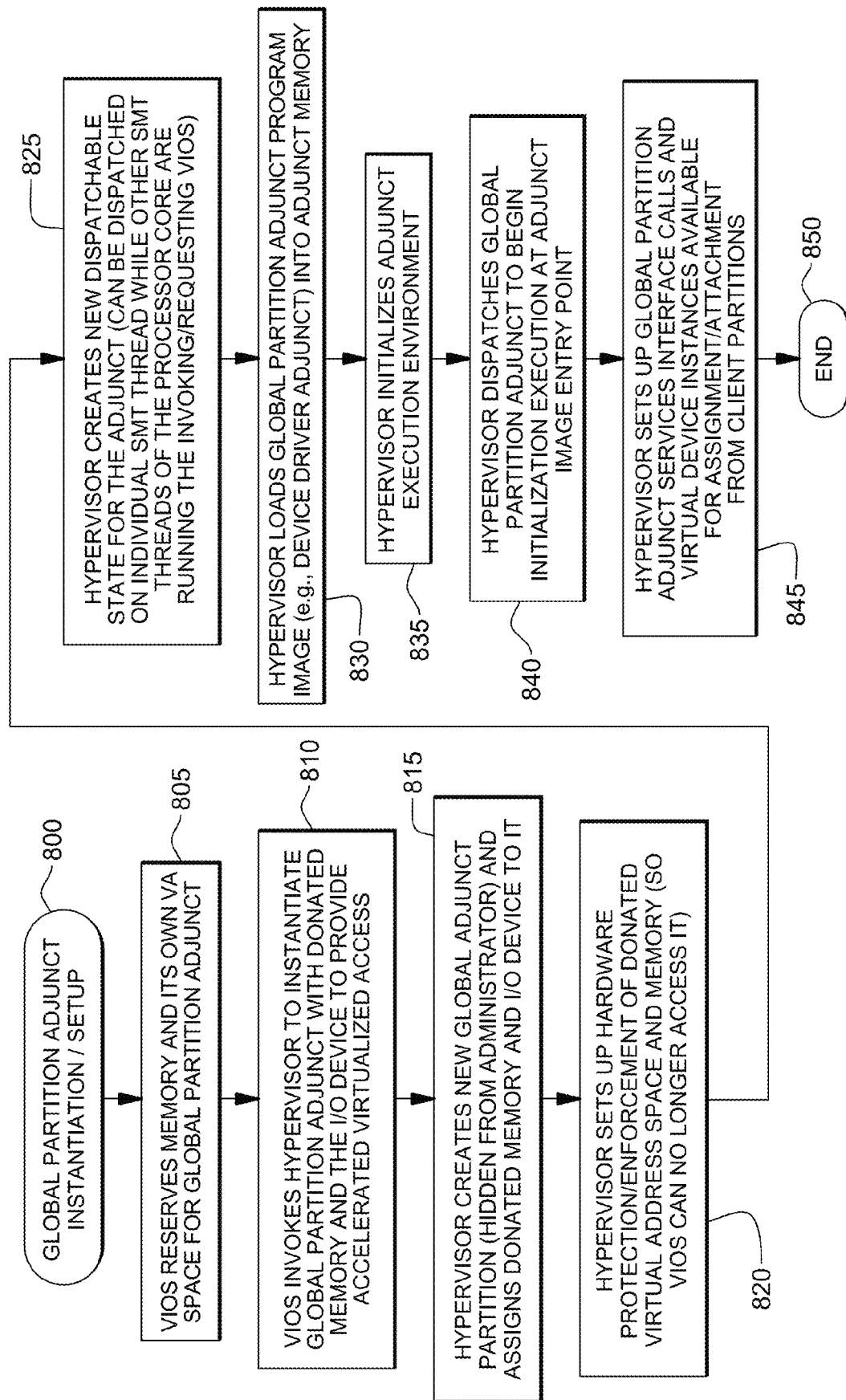


FIG. 8

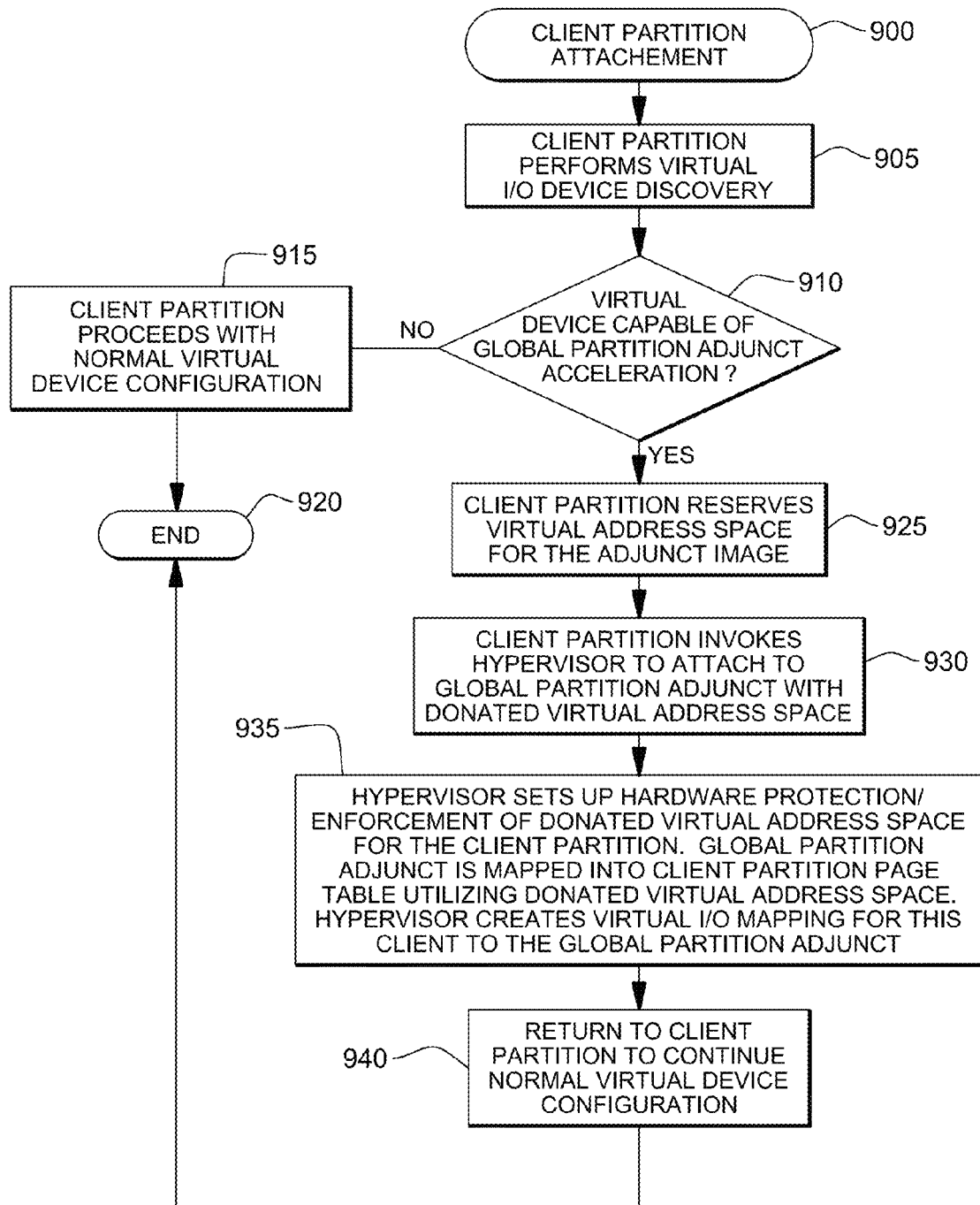


FIG. 9

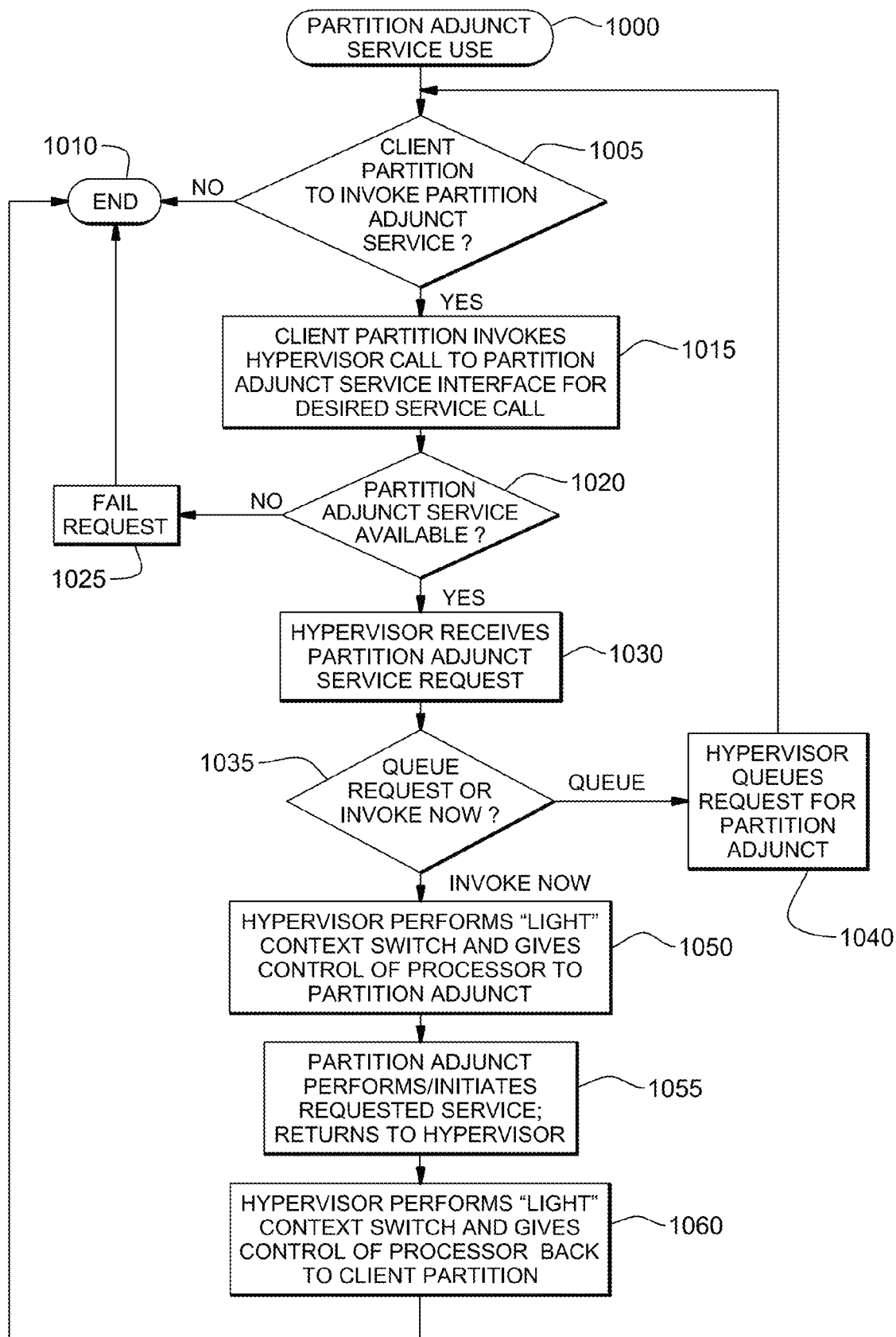


FIG. 10

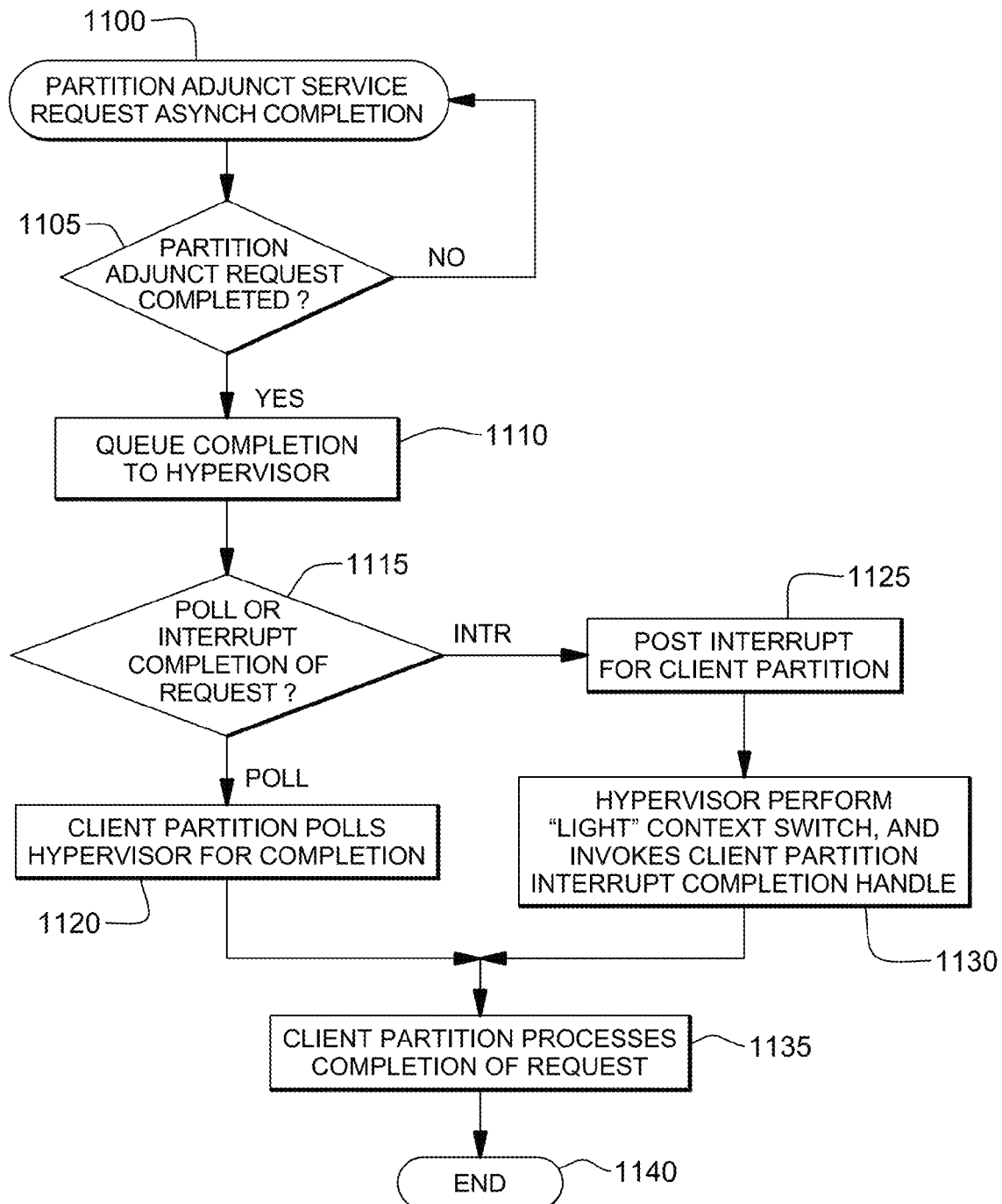


FIG. 11

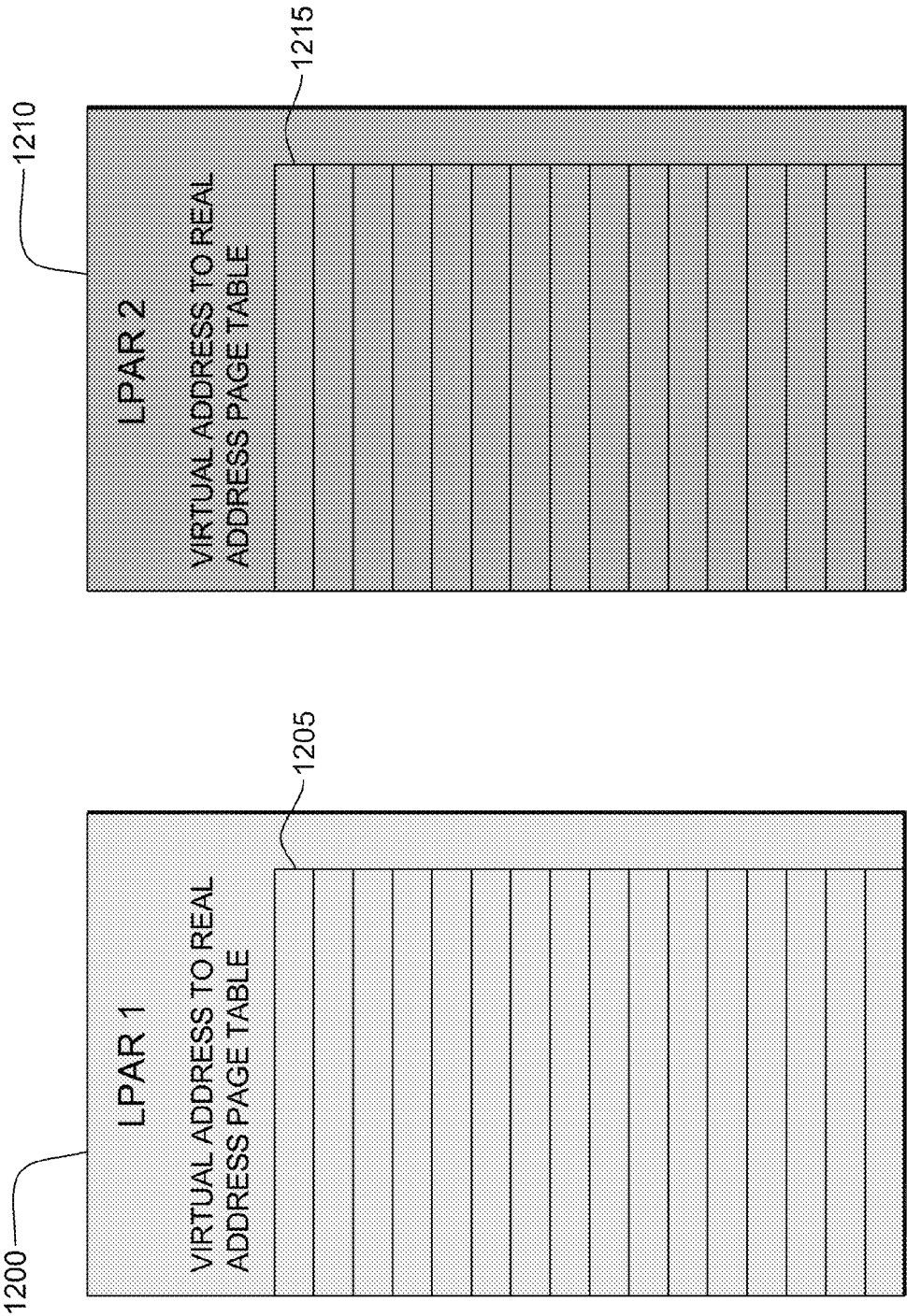
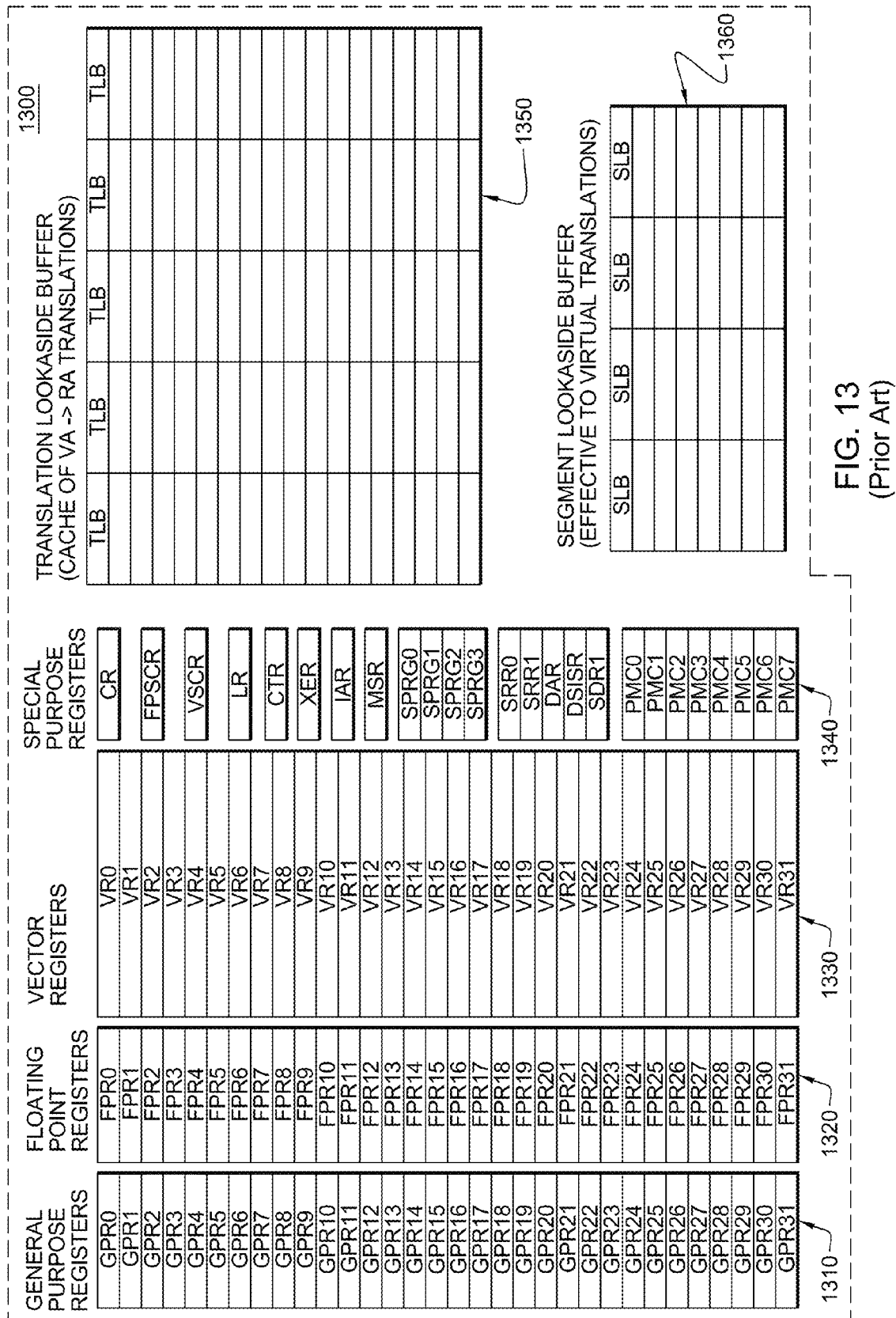


FIG. 12
(Prior Art)



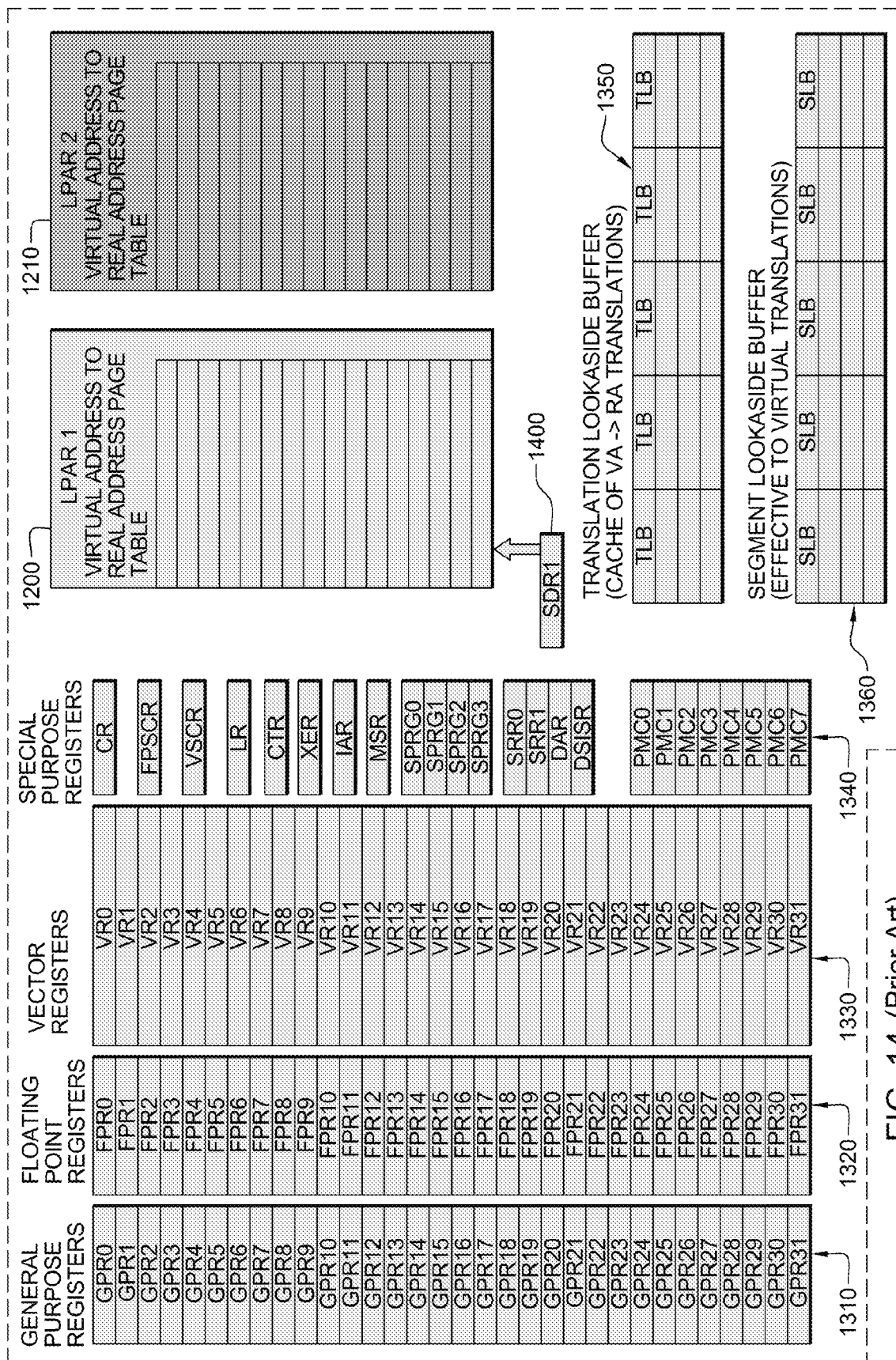


FIG. 14 (Prior Art)

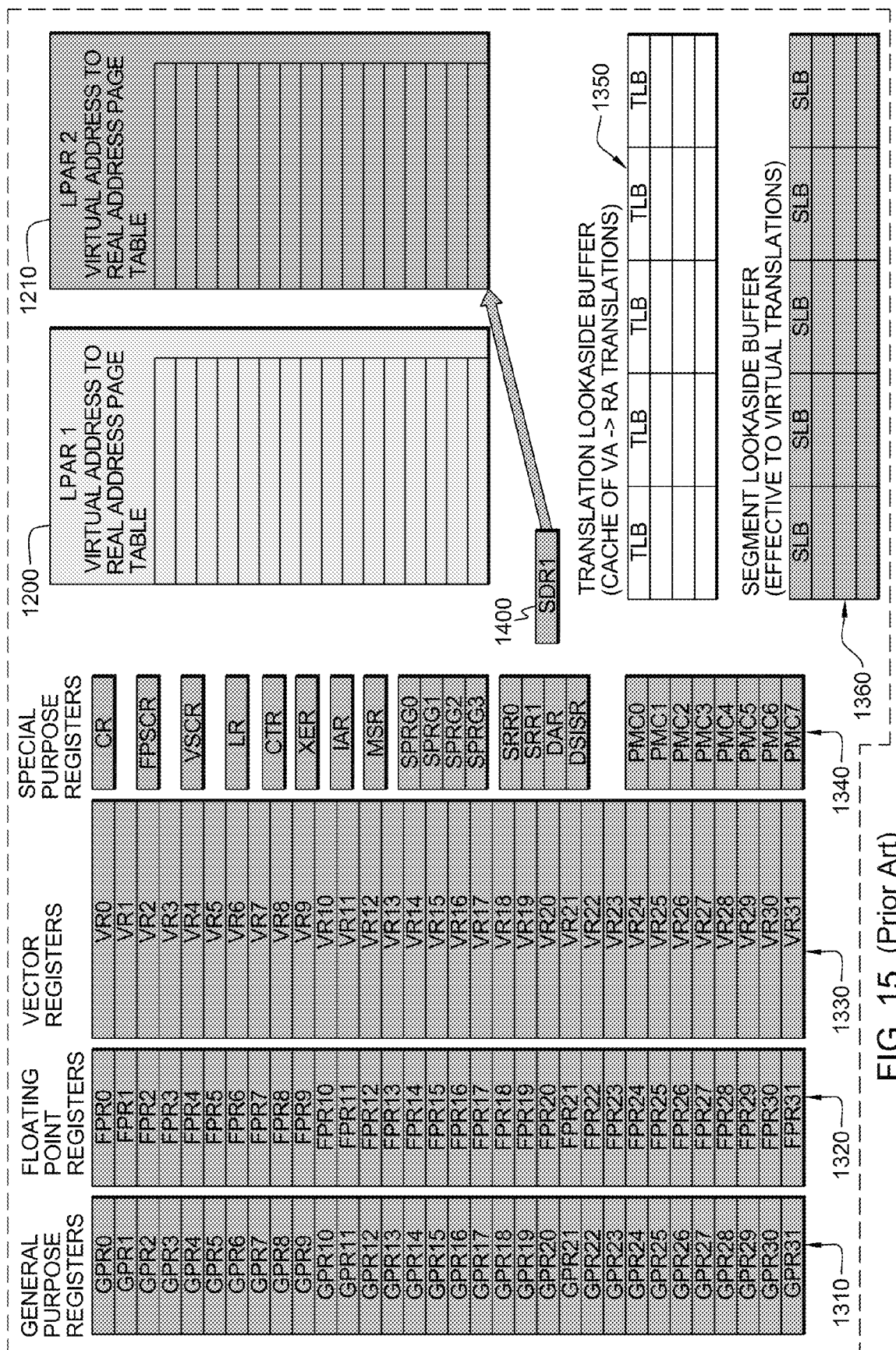


FIG. 15 (Prior Art)

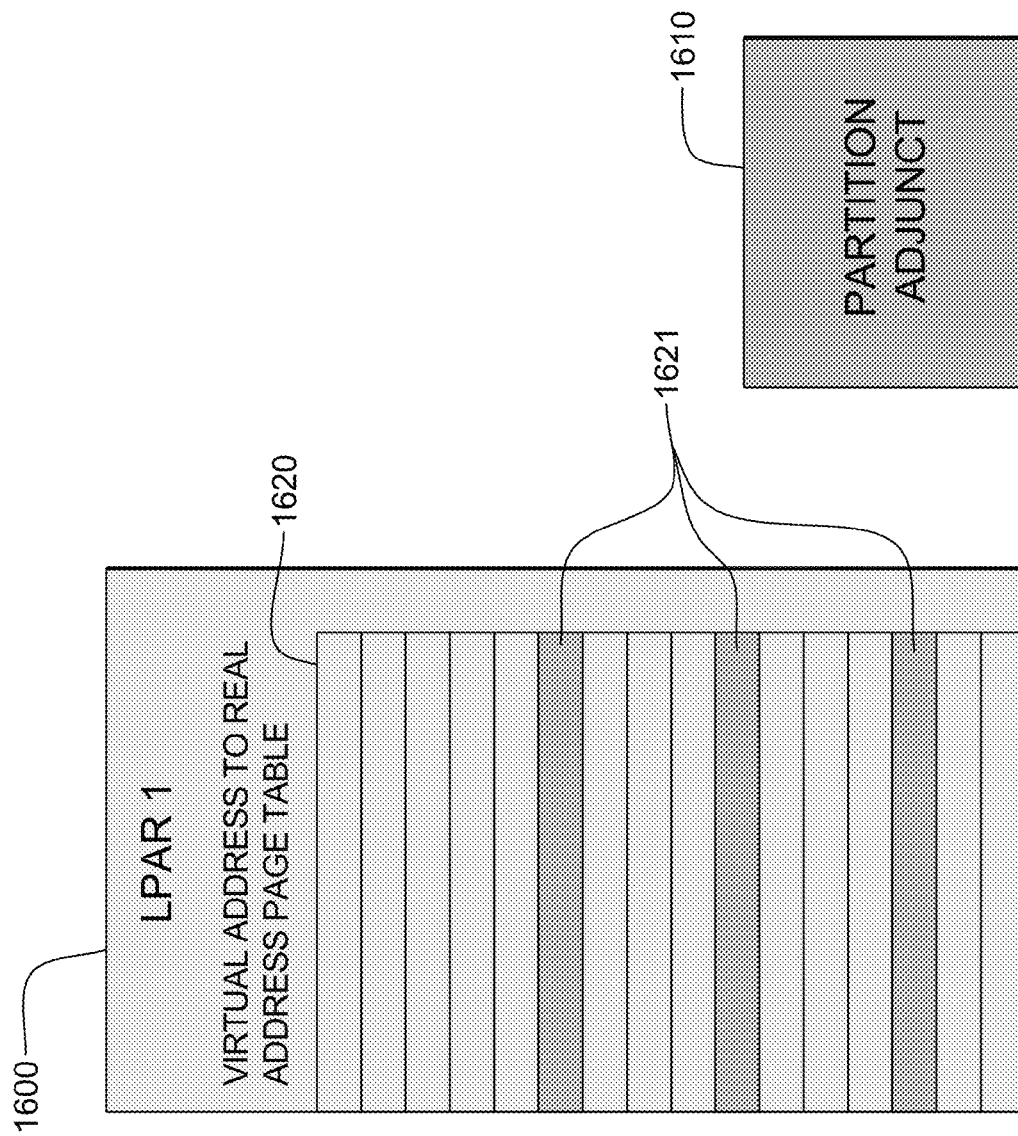


FIG. 16

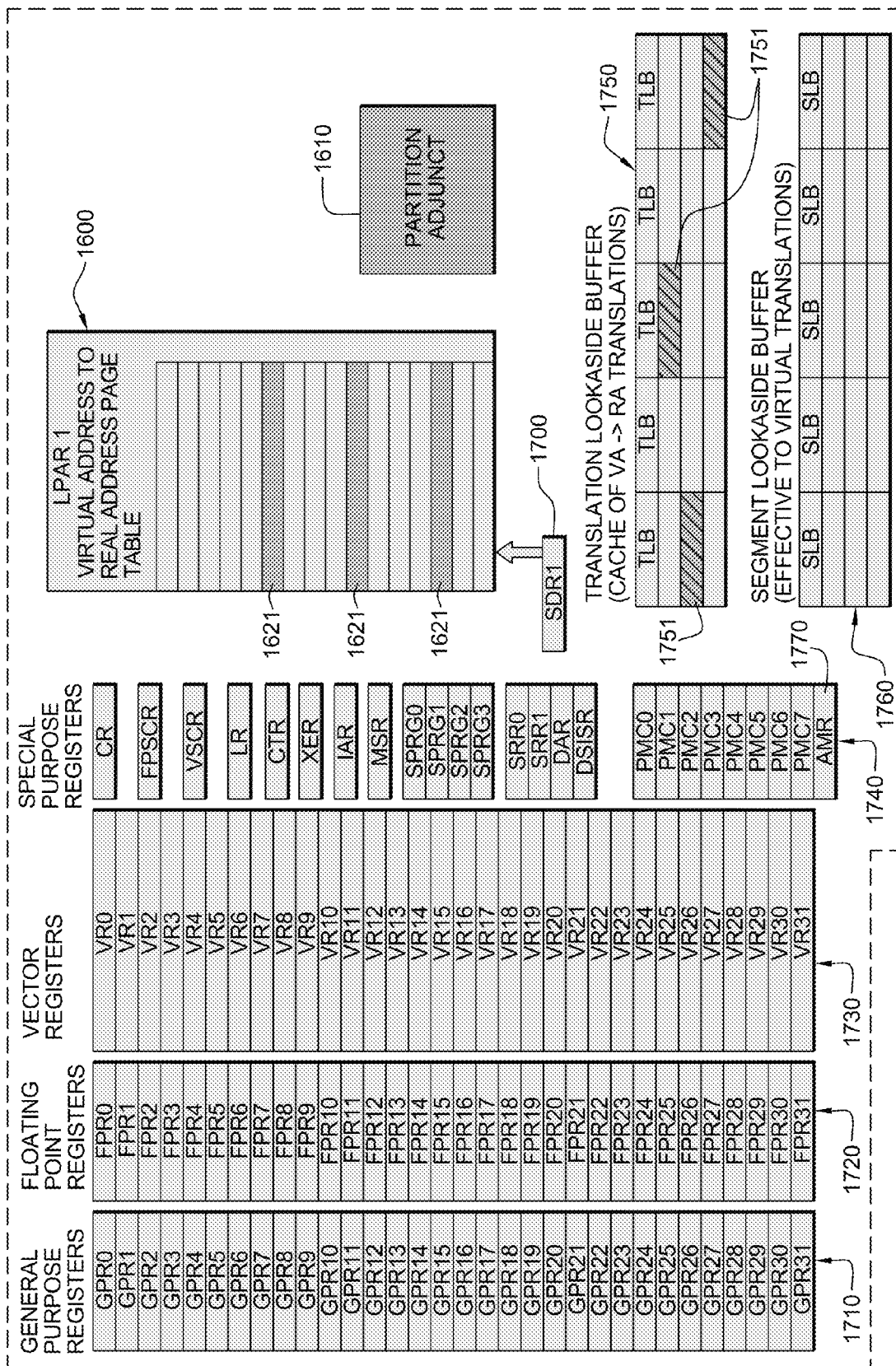


FIG. 17

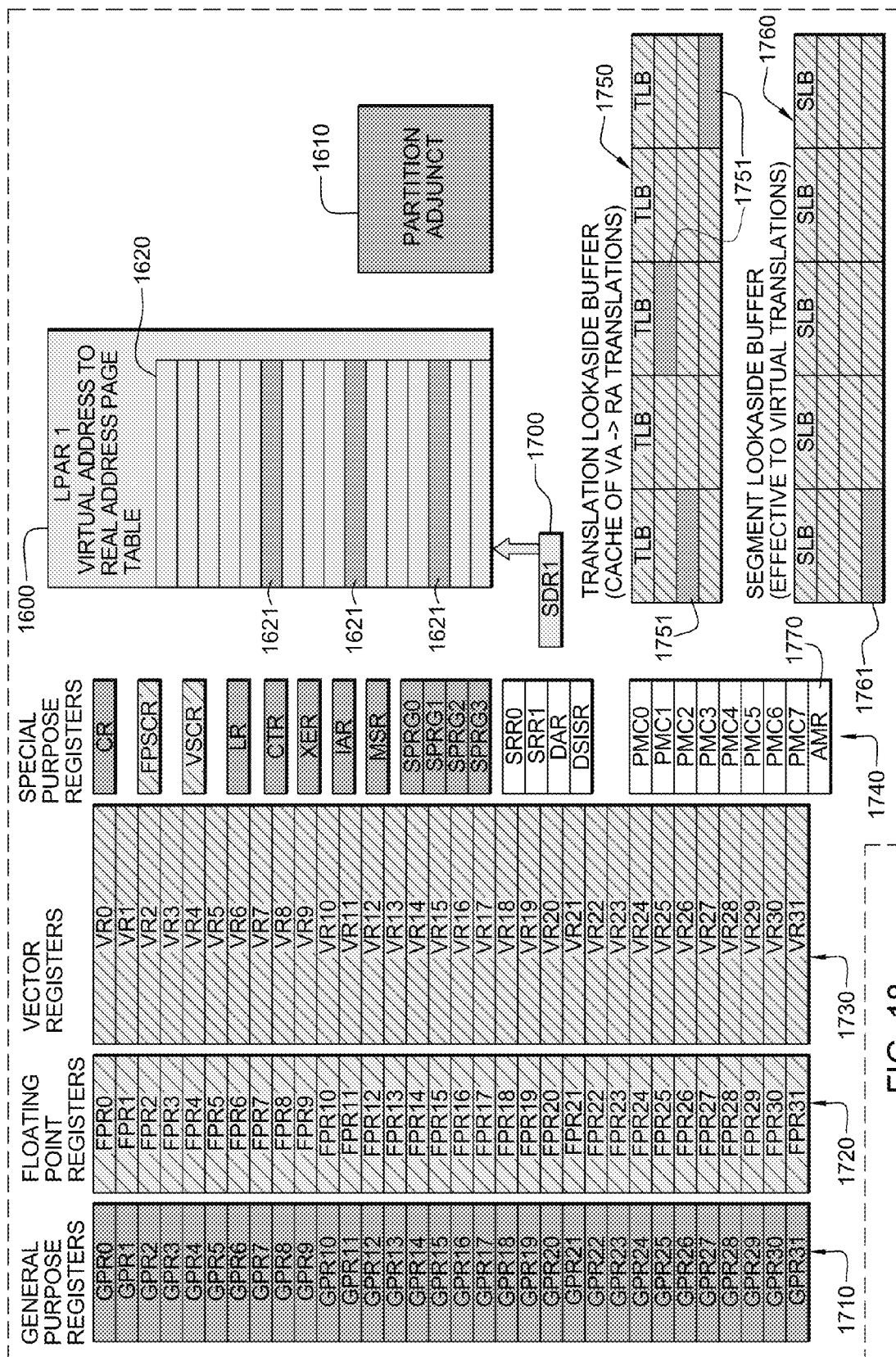
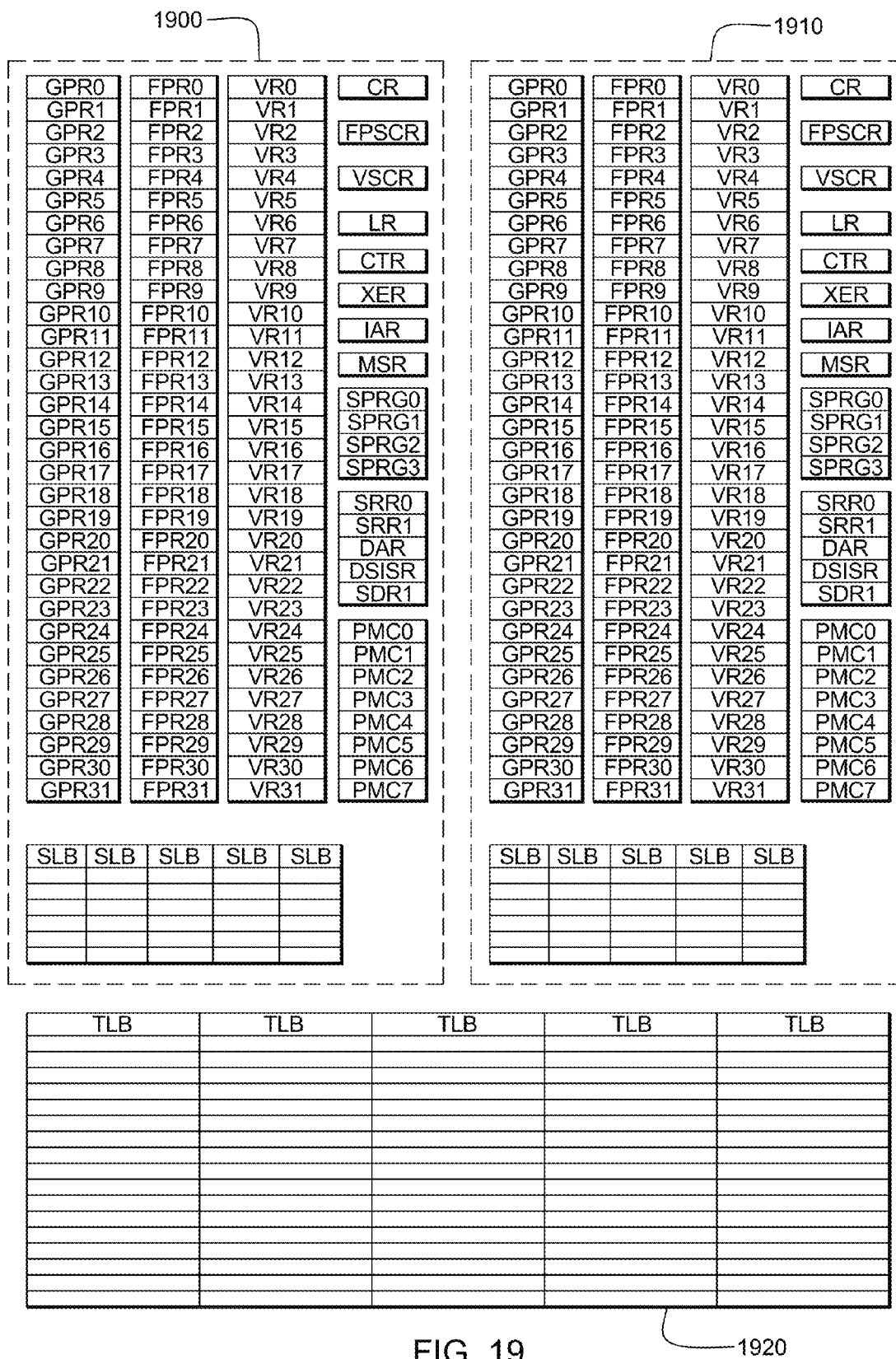
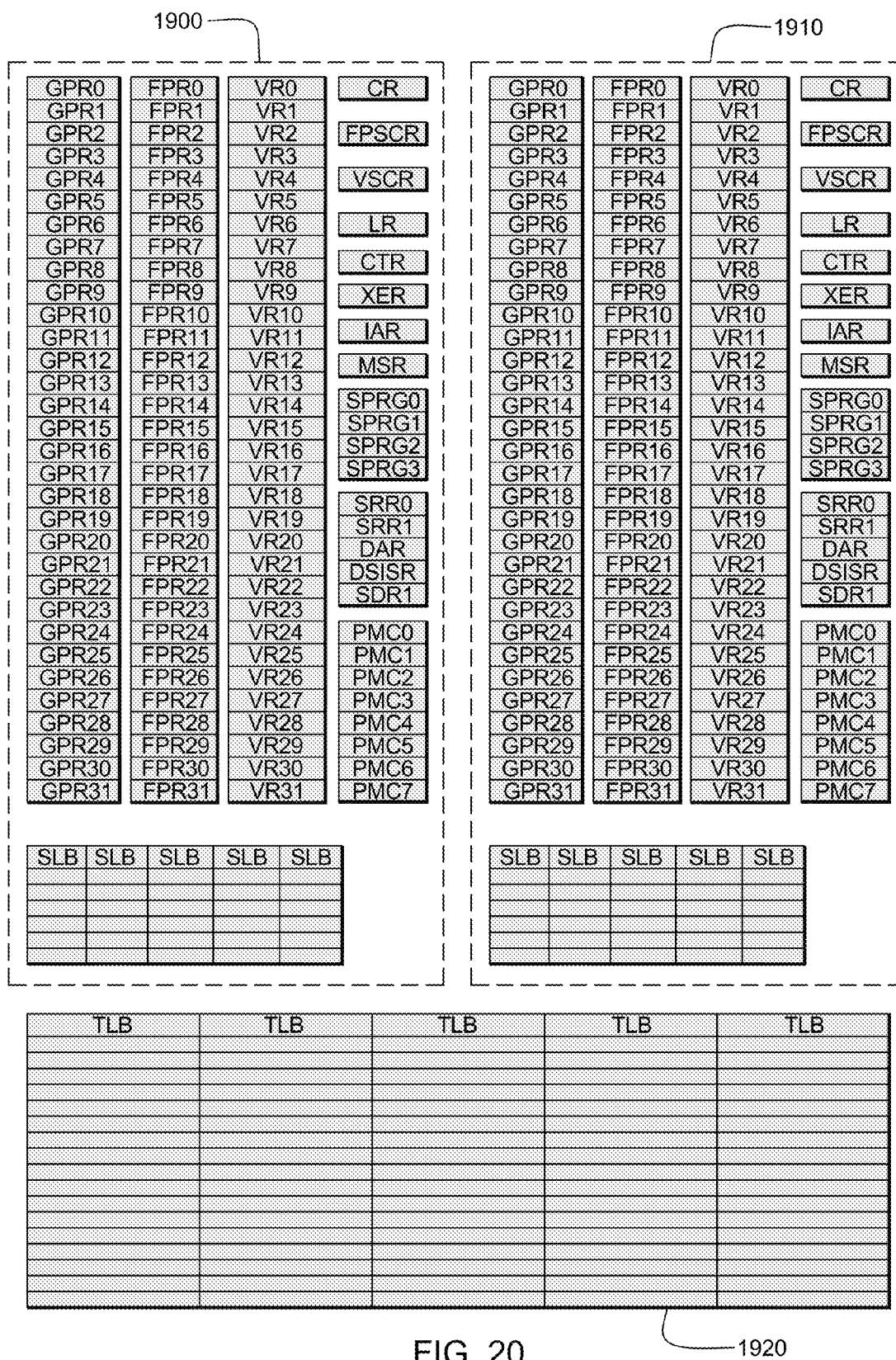
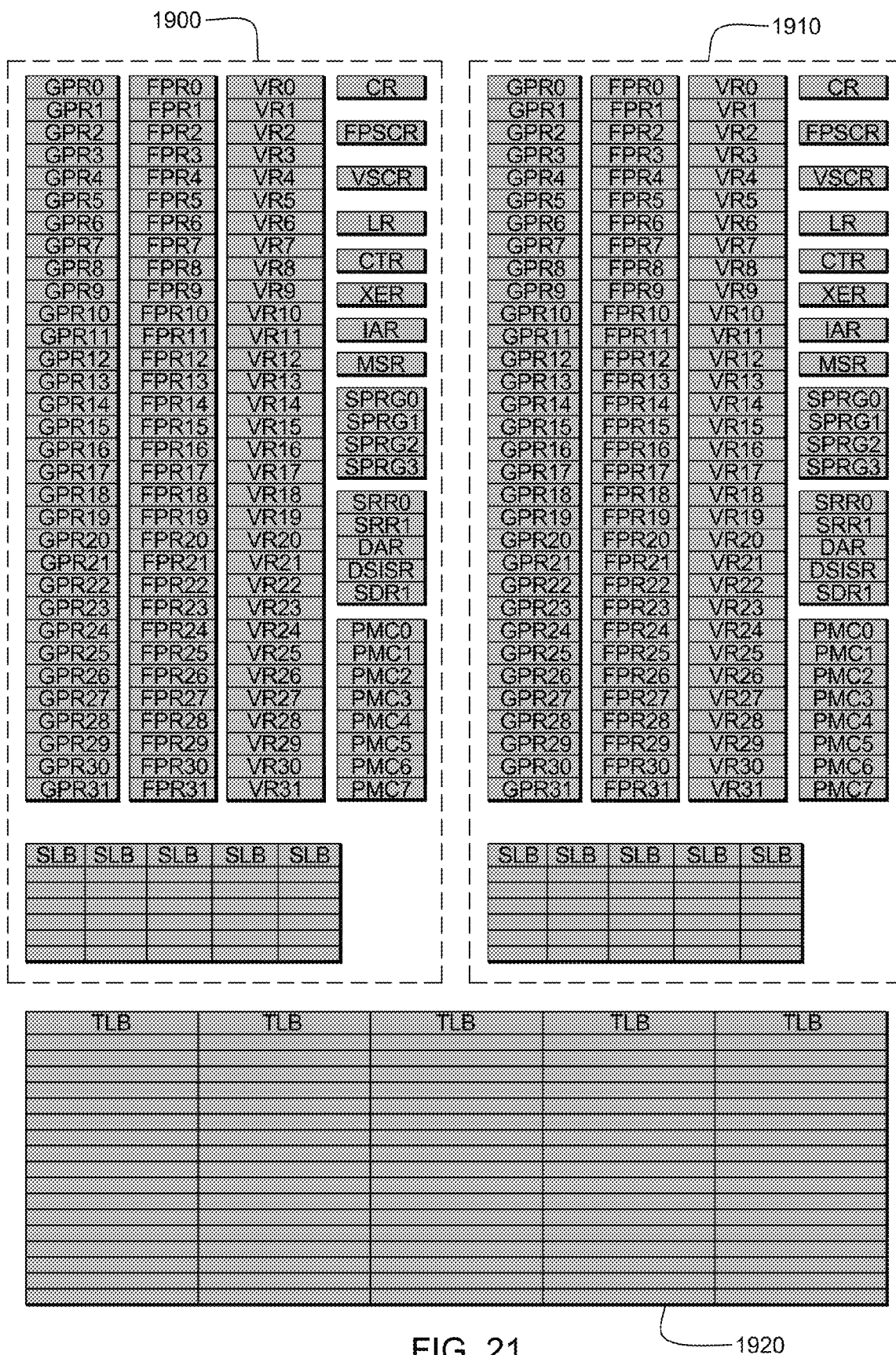
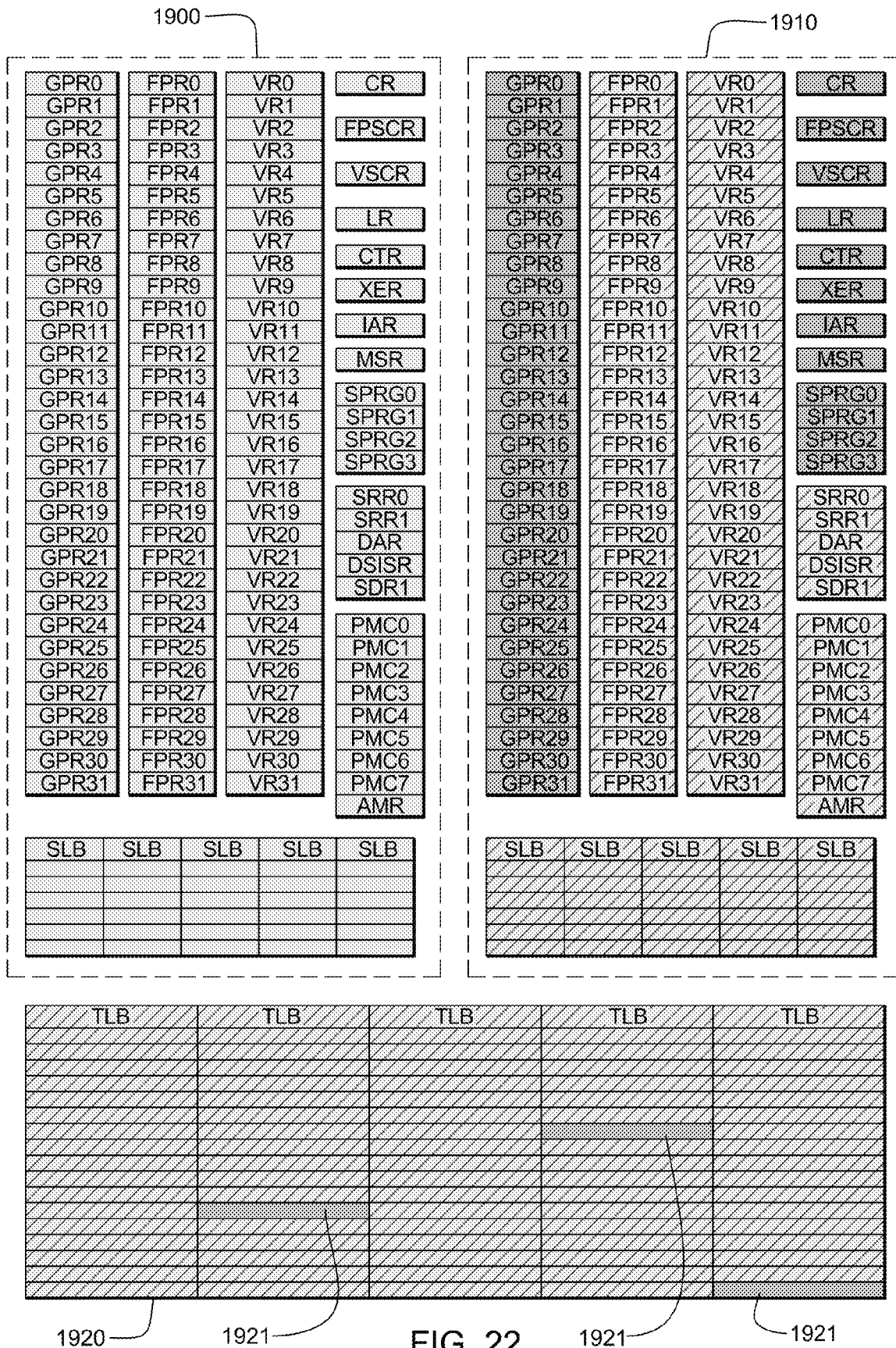


FIG. 18









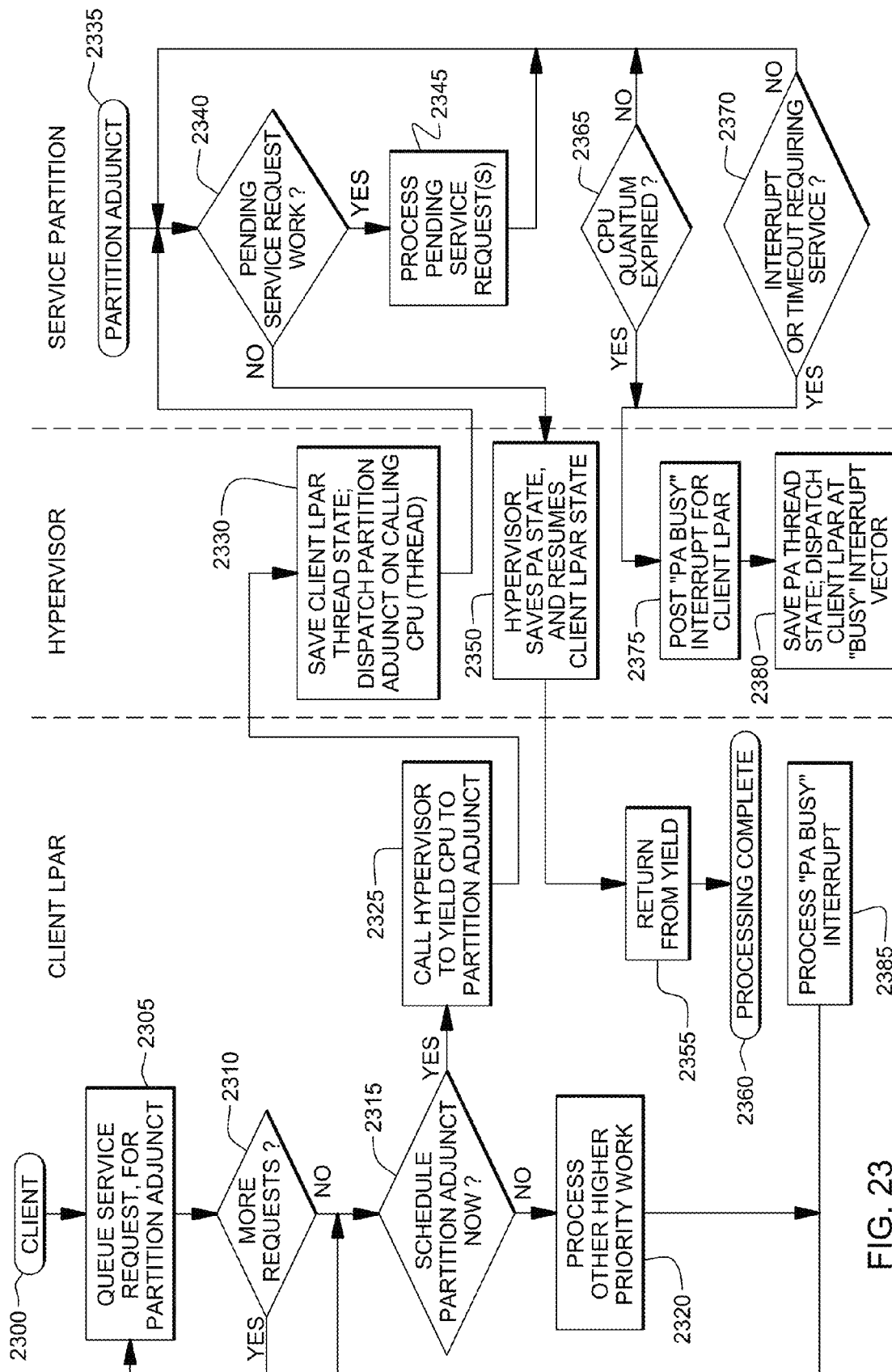


FIG. 23

COMPUTER
PROGRAM
PRODUCT
2400

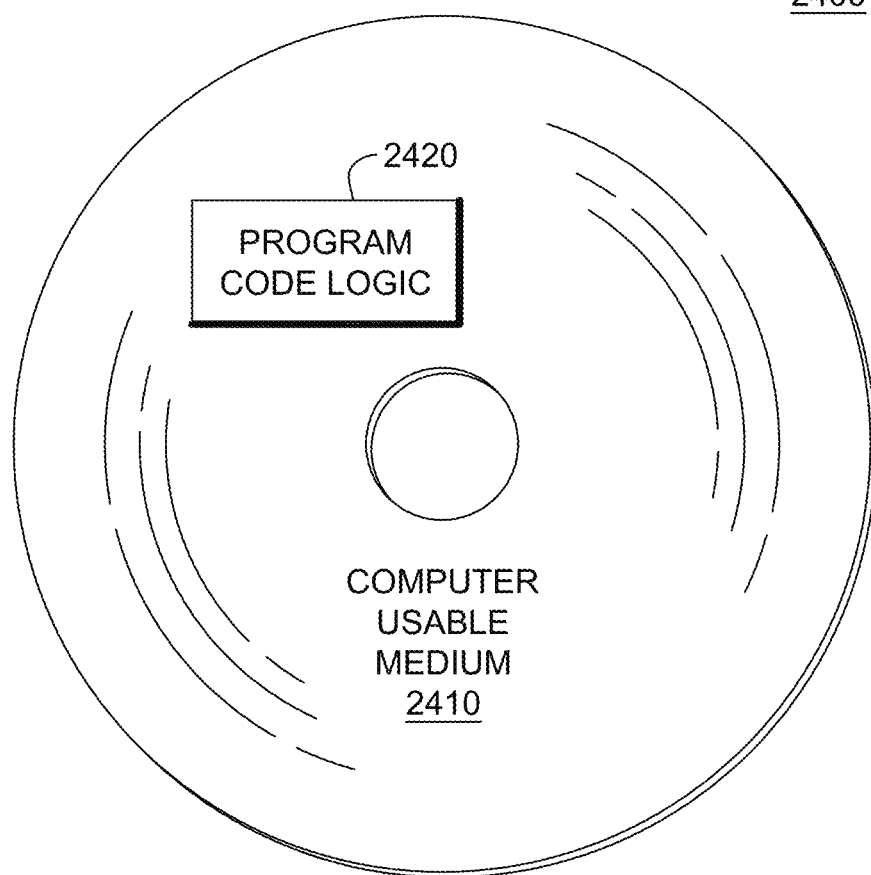


FIG. 24

1

CLIENT PARTITION SCHEDULING AND PRIORITIZATION OF SERVICE PARTITION WORK

CROSS-REFERENCE TO RELATED APPLICATION

This application is a continuation of U.S. patent application Ser. No. 12/110,953, filed Apr. 28, 2008, entitled "Client Partition Scheduling and Prioritization of Service Partition Work", which was published on Feb. 5, 2009, as U.S. Patent Publication No. 2009/0037907 A1, and which claims the benefit of U.S. provisional application Ser. No. 60/953,512, filed Aug. 2, 2007, entitled "Partition Adjunct Environment", the entirety of each of which is incorporated herein by reference.

TECHNICAL FIELD

The present invention relates generally to data processing systems, and more particularly, to client partition scheduling and prioritization of a service partition. In one implementation, the service partition is a partition adjunct comprising a new execution environment for the data processing system.

BACKGROUND OF THE INVENTION

Logical partitions (LPARs) running atop a hypervisor of a data processing system are often used to provide higher-level function than provided by the hypervisor itself. For example, one LPAR may be designated a virtual input/output server (VIOS), which provides input/output services to one or more other LPARs of the data processing system. This offloading of higher-level function avoids complex code in the hypervisor, and thus, assists in maintaining the hypervisor small and secure within the data processing system. Using logical partitions to provide such services to other logical partitions, however, requires relatively high overhead to instantiate and run the logical partition, and thus, a full operating system, in order to provide such services.

SUMMARY OF THE INVENTION

Disclosed herein is a solution which allows services to be provided to a logical partition without requiring another full logical partition to provide the services, and without requiring embedding of the services in the initiating logical partition or in the hypervisor. A partition adjunct, such as described and claimed herein, is a new execution environment which provides this middle ground.

Provided herein, in one aspect, is a method in a data processing system for processing a service request. The method includes: obtaining by a service partition of the data processing system a service request from a logical partition of the data processing system, the logical partition and the service partition each executing above a hypervisor of the data processing system; and processing the service request by the service partition utilizing a processor quantum assigned to the logical partition and donated by the logical partition to the service partition.

In another aspect, a data processing system is provided which includes at least one processing environment which comprises a logical partition and a service partition to process a service request for the logical partition. The logical partition and the service partition each execute above a hypervisor of the data processing system. The service request is processed by the service partition utilizing a processor quantum

2

assigned to the logical partition and donated by the logical partition to the service partition with a call to the service partition to process the service request.

In a further aspect, the invention comprises an article of manufacture which includes at least one computer-usable medium having computer-readable program code logic to define a method of processing a service request within a data processing system. The computer-readable program code logic when executing on a processor performing: obtaining by a service partition of the data processing system a service request from a logical partition of the data processing system, the logical partition and the service partition each executing above a hypervisor of the data processing system; and processing the service request by the service partition utilizing a processor quantum assigned to the logical partition and donated by the logical partition to the service partition.

Further, additional features and advantages are realized through the techniques of the present invention. Other embodiments and aspects of the invention are described in detail herein and are considered a part of the claimed invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The subject matter which is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other objects, features, and advantages of the invention are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

FIG. 1 is a block diagram of one embodiment of a data processing system to implement one or more aspects of the present invention;

FIG. 2 is a more detailed illustration of a data processing system which could be used to implement one or more aspects of the present invention;

FIG. 3 illustrates one embodiment of a data processing system comprising a client logical partition and a partition adjunct, in accordance with an aspect of the present invention;

FIG. 4A illustrates a data processing system comprising multiple logical partitions and multiple partition adjuncts, in accordance with an aspect of the present invention;

FIG. 4B illustrates a data processing system wherein multiple logical partitions attach a global partition adjunct, in accordance with an aspect of the present invention;

FIG. 4C depicts a data processing system wherein a hypervisor employs a partition adjunct, in accordance with an aspect of the present invention;

FIG. 5 illustrates a data processing system wherein a global partition adjunct, comprising a device driver service, is attached by multiple client partitions instead of employing a conventional virtual input output server, in accordance with an aspect of the present invention;

FIG. 6 is a flowchart of one embodiment of a process for creating a partition adjunct, in accordance with an aspect of the present invention;

FIG. 7 is a flowchart of a more detailed embodiment of a process for creating a partition adjunct comprising a device driver, in accordance with an aspect of the present invention;

FIG. 8 is a flowchart of one embodiment of a process for creating a global partition adjunct employable by multiple logical partitions of a data processing system, in accordance with an aspect of the present invention;

FIG. 9 is a flowchart of one embodiment of a process for attaching to a global partition adjunct from a client logical partition, in accordance with an aspect of the present invention;

3

FIG. 10 is a flowchart of one embodiment of a process for processing a service request by a partition adjunct, in accordance with an aspect of the present invention;

FIG. 11 is a flowchart of one embodiment of a process for asynchronous service request completion handling, in accordance with an aspect of the present invention;

FIG. 12 illustrates two conventional logical partitions of a data processing system;

FIG. 13 illustrates conventional processor register machine state and memory management and address translation hardware of a data processing system;

FIG. 14 illustrates the processor register machine state and memory management unit hardware of FIG. 13, with LPAR 1 conventionally dispatched in an active state;

FIG. 15 illustrates the processor register machine state and memory management unit hardware of FIG. 13, with LPAR 1 switched out, and LPAR 2 dispatched in the active state;

FIG. 16 illustrates a data processing system comprising a logical partition (LPAR 1) and a partition adjunct, in accordance with an aspect of the present invention;

FIG. 17 illustrates processor register machine state and memory management and address translation hardware of the data processing system of FIG. 16, with LPAR 1 dispatched in an active state, and illustrating shared virtual address space between LPAR 1 and the partition adjunct, in accordance with an aspect of the present invention;

FIG. 18 illustrates the processor register machine state and memory management and address translation hardware of FIG. 17 after context switching the current state machine from LPAR 1 to dispatch the partition adjunct, in accordance with an aspect of the present invention;

FIG. 19 illustrates processor register machine state and memory management and address translation hardware for a multithreaded processor core of a data processing system;

FIG. 20 illustrates the processor register machine state and memory management and address translation hardware of FIG. 19, with a first logical partition (LPAR 1) in the active state;

FIG. 21 illustrates the processor register machine state and memory management and address translation hardware of FIGS. 19 & 20 after context switching the current state machine to state data of a second logical partition (LPAR 2) from LPAR 1;

FIG. 22 illustrates the processor register machine state and memory management and address translation hardware of FIG. 19, with partial context switching of the current state machine of one hardware thread from a first logical partition (LPAR 1) to the partition adjunct, in accordance with an aspect of the present invention;

FIG. 23 is a flowchart of one embodiment of a process for client partition control of prioritizing and scheduling service requests for execution by a service partition of a data processing system, in accordance with an aspect of the present invention; and

FIG. 24 depicts one embodiment of a computer program product incorporating one or more aspects of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

FIG. 1 is a block diagram of a data processing system 100, which in one example, is a symmetric multiprocessing (SMP) server computer system. SMP server computer system 100 includes physical hardware devices that can be mapped to, i.e., temporarily owned by, a user application to execute that application.

4

SMP server computer system 100 includes a physical SMP server 102. Physical SMP server 102 includes physical hardware devices such as processor 104, memory 106, and I/O adapters 108. These physical devices are managed by hypervisor 110. Processors 104 are shared processors and each may be a simultaneous multithreading (SMT)-capable processor that is capable of concurrently executing multiple different threads on the processor.

A virtual server is a proxy for a physical server that has the same capabilities, interfaces, and state. Virtual servers are created and managed by a hypervisor that resides on physical SMP server computer system 100. A virtual server appears to be a physical SMP server to its user: the operating system, middleware, and application software that run upon it. SMP server computer system 100 includes one or more virtual servers such as virtual server 112 and virtual server 112a.

Each virtual server appears to its software to include its own processor(s), memory, and I/O adapter(s) that are available for the exclusive use of that virtual server. For example, virtual server 112 includes a virtual processor 120, virtual memory 122, and virtual I/O adapters 124. Virtual server 112a includes virtual processors 120a, virtual memory 122a, and virtual I/O adapters 124a.

Each virtual server supports its own software environment, including an operating system, middleware, and applications. The software environment of each virtual server can be different from the software environment of other virtual servers. For example, the operating systems executed by each virtual server may differ from one another.

For example, virtual server 112 supports operating system 114, middleware 116, and applications 118. Virtual server 112a supports operating system 114a, middleware 116a, and applications 118a. Operating systems 114 and 114a may be the same or different operating systems.

A virtual server is a logical description of a server that defines a server environment that acts, to a user, as if it were a physical server, being accessed and providing information in the same way as a physical server. The virtual processors, virtual memory, and virtual I/O adapters that are defined for each virtual server are logical substitutes for physical processors, memory, and I/O adapters.

Hypervisor 110 manages the mapping between the virtual servers with their virtual processors, virtual memory, and virtual I/O adapters and the physical hardware devices that are selected to implement these virtual devices. For example, when a virtual processor is dispatched, a physical processor, such as one of physical processors 104, is selected by hypervisor 110 to be used to execute and implement that virtual processor. Hypervisor 110 manages the selections of physical devices and their temporary assignment to virtual devices.

Hypervisor 110 services all of the logical partitions during a dispatch time slice. The dispatch time slice is a particular length of time. During each dispatch time slice, hypervisor 110 will allocate, or assign, the physical processor to each logical partition. When the logical partition has been allocated time on the physical processor, the virtual processors defined by that logical partition will be executed by the physical processor.

Hypervisor 110 is responsible for dynamically creating, managing, and destroying virtual SMP servers. Whole virtual processors, virtual I/O adapters, and virtual memory blocks can be removed or added by hypervisor 110. Hypervisor 110 is also responsible for dynamic resource allocation, managing time-sharing of physical resources, and altering the physical resource mapped to a processor without involving the operating system. Hypervisor 110 is also able to dedicate physical resources to virtual resources for situations where

sharing is not desired. Hypervisor **110** is responsible for managing the addition or removal of physical resources. Hypervisor **110** makes these additions and deletions transparent to the upper level applications.

FIG. 2 is a more detailed illustration of a computer system that may be used to implement the concepts described herein. Data processing system **200** may be a symmetric multiprocessor (SMP) system including a plurality of shared processors or SMT-capable processors, such as processors **202** and **204** connected to system bus **206**. Alternatively, a single processor system may be employed. In the depicted example, processor **204** is a service processor. Each SMT-capable processor is capable of concurrently executing multiple hardware threads on the one processor.

Also connected to system bus **206** is memory controller/cache **208**, which provides an interface to local memory **209**. I/O bus bridge **210** is connected to system bus **206** and provides an interface to I/O bus **212**. Memory controller/cache **208** and I/O bus bridge **210** may be integrated as depicted.

Peripheral component interconnect (PCI) bus bridge **214** connected to I/O bus **212** provides an interface to PCI local bus **216**. A number of modems may be connected to PCI bus **216**. Typical PCI bus implementations will support four PCI expansion slots or add-in connectors. Communications links to network computers **108-112** in FIG. 1 may be provided through modem **218** and network adapter **220** connected to PCI local bus **216** through add-in boards.

Network adapter **220** includes a physical layer **282** which conditions analog signals to go out to the network, such as for example, an Ethernet network for an R45 connector. A media access controller (MAC) **280** is included within network adapter **220**. Media access controller (MAC) **280** is coupled to bus **216** and processes digital network signals. MAC **280** serves as an interface between bus **216** and physical layer **282**. MAC **280** performs a number of functions involved in the transmission and reception of data packets. For example, during the transmission of data, MAC **280** assembles the data to be transmitted into a packet with address and error detection fields. Conversely, during the reception of a packet, MAC **280** disassembles the packet and performs address checking and error detection. In addition, MAC **280** typically performs encoding/decoding of digital signals transmitted and performs preamble generation/removal as well as bit transmission/reception.

Additional PCI bus bridges **222** and **224** provide interfaces for additional PCI buses **226** and **228**, from which additional modems or network adapters may be supported. In this manner, data processing system **200** allows connections to multiple network computers. A memory-mapped graphics adapter **230** and hard disk **232** may also be connected to I/O bus **212** as depicted, either directly or indirectly.

Service processor **204** interrogates system processors, memory components, and I/O bridges to generate and inventory and topology understanding of data processing system **200**. Service processor **204** also executes Built-In-Self-Tests (BISTs), Basic Assurance Tests (BATs), and memory tests on all elements found by interrogating a system processor, memory controller, and I/O bridge. Any error information for failures detected during the BISTs, BATs, and memory tests are gathered and reported by service processor **204**.

Those of ordinary skill in the art will appreciate that the hardware depicted in FIG. 2 may vary. For example, other peripheral devices, such as optical disk drives and the like, also may be used in addition to or in place of the hardware depicted. The depicted example is not meant to imply architectural limitations with respect to the present invention.

The present invention may be executed within one of the computers or data processing systems depicted in FIG. 1 or 2. As a specific, commercially available example, the data processing system implementing a partition adjunct such as described hereinbelow can be built upon technologies found in IBM's p/i Series product line firmware and systemware, as described in the "Power Architecture Platform Reference" (PAPR) material at Power.org (http://www.power.org/members/developers/specs/PAPR_Version_2.7_09Oct07.pdf), which is hereby incorporated by reference.

Disclosed hereinbelow is the concept of a partition adjunct, which is a partition that is more limited than a full, logical partition. As described below, the partition adjunct runs in a flat, static effective address space and problem state. These restrictions permit the hypervisor to apply a range of hypervisor and processor optimizations that result in a substantial decrease in system overhead associated with a context switch of the state machine from a logical partition to state data of a partition adjunct, that is, compared to a context switch of the state machine between logical partitions. In other respects, a partition adjunct is similar to a full logical partition. For example, a partition adjunct can be assigned resources, either physical or virtual, similar to a full logical partition. Further, a partition adjunct can be an end-point of a virtual input output (VIO) communications mechanism, similar to a full logical partition, such as a virtual input output server (VIOS).

FIG. 3 illustrates one example of a data processing system **300**, comprising a client logical partition (or simply client partition) **310** and a partition adjunct **320**, in accordance with an aspect of the present invention. Client partition **310** is a logical partition which comprises one or more customer applications **312**, an operating system instance **314** and a virtual I/O facility **316**. Client partition **310** runs atop a hypervisor **330**, and provides higher-level function than that provided by the hypervisor itself. Hypervisor **330** executes atop the underlying system hardware **340**, which is shown to include one or more I/O devices **350**.

Partition adjunct **320** is conceptually a child partition to client partition **310**. The partition adjunct is less than a full logical partition, but is run in a manner whereby the hypervisor enforces security and isolation between the partition adjunct and the client partition it runs within. The partition adjunct is provided with reduced functionality compared with a full logical partition, for example, has no access to memory management unit (MMU) configuration or floating-point facilities, and is an environment that only the functionality needed to run the desired service (e.g., I/O driver) is provided. In the illustrated example, partition adjunct **320** includes a virtual I/O interface **322** and a hardware device driver service **324**, which allows access to I/O device **350**. In operation, client partition **310** accesses I/O device **350** via the partition adjunct **320**, as illustrated. By reducing functionality within the partition adjunct environment, the run time overhead of dispatching and maintaining the partition adjunct (compared with another full logical partition) is reduced, and consequently, many of the performance disadvantages of using a separate logical partition as a virtual input output server (VIOS) are avoided.

As a specific example, the partition adjunct is described herein as running a reduced operating system environment for a device driver service. This service is provided by way of example only. The partition adjunct provides minimal, optimized infrastructure consisting only of that needed by a device driver (in one example). For instance, if a Linux device driver were run inside of the partition adjunct, the minimal execution environment would consist of the Linux kernel services, or equivalent services, that the Linux device driver

required. If an AIX device driver were run inside of the partition adjunct, then the minimal execution environment would consist of the AIX kernel services, or equivalent services, that the AIX device driver required. Advantageously, the partition adjunct runs in hypervisor/problem-state, directly against hypervisor interfaces. As explained in detail below, dispatching of the partition adjunct does not require a full partition context switch, which simplifies adjunct kernel requirements. This is achieved, in part, by mapping the partition adjunct into the client partition's virtual address page table. Client partition to partition adjunct isolation can be achieved, for example, via hypervisor-managed memory keys. Advantageously, the partition adjunct is not customer viewable. Further, the same partition adjunct service (referred to herein as a global partition adjunct service) may be instantiated within multiple client partitions, as explained below.

FIGS. 4A-4C depict various partition adjunct usage models which can be implemented in accordance with the concepts disclosed herein for addressing a number of operating system and platform issues. FIG. 4A illustrates one example of a partition adjunct 420 (also referred to in this example as a local partition adjunct), which conceptually, partially resides within an initiating client partition 410 for accessing dedicated resources 440 through a hypervisor 430. For example, a common adapter/driver service may be provided by partition adjunct 420 for a respective dedicated adapter (i.e., resource).

FIG. 4B illustrates an example of a global partition adjunct, wherein, (for example) a service logical partition such as a virtual input output server partition, donates memory and physical resources for instantiation of the partition adjunct. Such a global partition adjunct 421 is accessible or attachable by multiple client partitions 410, and provides, for example, input output services to a resource 440 via hypervisor 430. As a specific example, the global partition adjunct may comprise a common adapter driver service, and the resource a shared adapter.

FIG. 4C is an alternate embodiment of a global partition adjunct 422, wherein hypervisor 430 provides resources for the adjunct. In this implementation, the hypervisor employs the partition adjunct for its own use, for example, for protection or isolation services that would otherwise exist in the hypervisor's execution domain.

Unless otherwise specified, the partition adjunct embodiment described hereinbelow is assumed to be an execution environment that is part of a client partition, yet separated from the client partition by hypervisor isolation and security mechanisms, such as referenced below and described in greater detail in commonly assigned, co-filed application entitled "Hypervisor-Enforced Isolation of Entities Within a Single Logical Partition's Virtual Address Space", Ser. No. 12/111,041, filed Apr. 28, 2008, and issued Aug. 30, 2011, as U.S. Pat. No. 8,010,763 B2, the entirety of which is hereby incorporated herein by reference. Further details of partition adjunct processing and usage are described in greater detail in commonly assigned, co-filed patent applications entitled "Partition Adjunct for Data Processing System", Ser. No. 12/110,923, filed Apr. 28, 2008, and published Feb. 5, 2009, as U.S. Patent Publication No. 2009/0037906 A1; "Partition Adjunct with Non-Native Device Driver for Facilitating Access to a Physical Input/Output Device", Ser. No. 12/110,982, filed Apr. 28, 2008, and published Feb. 5, 2009, as U.S. Patent Publication No. 2009/0037908 A1; and "Multiple Partition Adjunct Instances Interfacing Multiple Logical Partitions to a Self-Virtualizing Input/Output Device", Ser. No. 12/111,020, filed Apr. 28, 2008, and published Feb. 5, 2009,

as U.S. Patent Publication No. 2009/0037941, the entirety of each of which is also hereby incorporated herein by reference.

FIG. 5 illustrates one detailed example of a global partition adjunct being employed by multiple client partitions, that is, the embodiment of FIG. 4B. In this example, logical partitions (LPAR A, LPAR B) 510A, 510B are client partitions, and partition 510C is a virtual input output server partition employed by client partitions 510A, 510B in, for example, providing an input/output service. In accordance with the present invention, the input/output service (referred to as the adjunct device driver 520 in this example), is a partition adjunct provided device driver service 520A, 520B for accessing input/output hardware 540 via one or more input/output paths 521 through hypervisor 530. In this example, the partition adjunct facilitates a more direct attachment or more direct path from each client partition to the input/output hardware. As described further below, the partition adjunct is incorporated into the virtual address spaces respectively donated by the client partitions employing the partition adjunct instances.

In order for a partition adjunct to be a runnable program, the hypervisor, along with a client partition that is to use the partition adjunct service, negotiate to establish the partition adjunct environment. Once this negotiation is complete, the client partition will have donated a portion of its virtual address space to the hypervisor for use by the partition adjunct. The hypervisor will use hardware and hypervisor facilities to ensure that the client partition no longer has access to or can modify the donated resources (e.g., the donated virtual address space). The hypervisor instantiates the effective address mappings required to run the partition adjunct using the donated virtual address resources. Subsequently, the hypervisor may switch between dispatching the client partition or the partition adjunct by reprogramming its control of the donated virtual address space. When the client partition runs, it may access all virtual address space assigned to it, except for the donated virtual address range, and when the partition adjunct runs, the hypervisor disables access to all virtual addresses of the client partition, except for the donated virtual address range, that is, the virtual address space to which it is enabled. This toggling of active/inactive virtual address ranges is significantly faster than reprogramming the full memory management and address translation hardware to effect a complete context switch of the current state machine between two full logical partitions, as is necessary to switch, for example, to a virtual input/output server partition. In this manner, the partition adjunct address space is carved out of and separated from the memory management and address translation hardware resources of the client partition. The partition adjunct is thus, from a processor's perspective, part of the client partition, but from the client partition's and hypervisor's perspective, is a distinct entity.

Advantageously, the partition adjunct concepts presented herein reduce the need to use full logical partitions for providing services to client partitions. This in turn frees up resources and improves performance for customer workloads. Additionally, the partition adjunct disclosed herein encourages the development and deployment of virtual platform services in lieu of development of operating system specific services by reducing the performance penalties associated with virtualized services. This in turn allows for savings and cost development, since services may be implemented only once (i.e., in a partition adjunct), rather than natively among multiple operating systems.

Various examples of protocol for instantiating and employing a partition adjunct are described below with reference to FIGS. 6-11.

FIG. 6 is a flowchart of one embodiment of a process for creating a local partition adjunct, for use such as depicted in FIG. 4A. The partition adjunct instantiation or setup protocol 600 begins with inquiry into whether the client partition wants a partition adjunct service 605. If “no”, then the protocol ends 610. Otherwise, the client partition reserves a portion of its virtual address space and memory space for the partition adjunct 615 and invokes the hypervisor to instantiate the partition adjunct with these donated resources 620. The hypervisor creates the new partition adjunct (which is hidden from the administrator), and assigns the donated resources to the partition adjunct 625. Creating the new partition adjunct means that the hypervisor creates a dispatchable program context. The hypervisor sets up hardware protection and enforcement of the donated virtual address space and donated memory, so that the client partition can no longer access that space 630. Further, the hypervisor creates a new dispatchable state for the partition adjunct 635. This new dispatchable state means that the partition adjunct can be dispatched on an individual hardware (SMT) thread of a multithreaded hardware core, while other SMT threads of the processor core continue to run the client partition invoking the partition adjunct. This aspect of the present invention is described further below with reference to FIGS. 19-22.

Next, the hypervisor loads the partition adjunct program image into the donated adjunct memory 640 and initializes adjunct execution 645. Initializing a partition adjunct execution environment can include the hypervisor initializing effective address space to virtual address translation tables, adjunct protection keys, and special hypervisor problem state privileges. The hypervisor then dispatches the partition adjunct to begin execution at the partition adjunct program entry point 650, and the hypervisor sets up partition adjunct service interface calls for the client partition 655. The hypervisor calls or invokes the client partition per protocols or functions provided by the partition adjunct. The hypervisor then returns to the client partition with status of the partition adjunct instantiation 660.

FIG. 7 depicts one detailed example of a process for creating a local partition adjunct, in accordance with an aspect of the present invention. In this example, the partition adjunct comprises a device driver service. Input/output instantiation or setup 700 begins with a client partition performing input/output device discovery 705. Processing inquires whether the client partition has its own device driver for a discovered input/output device 710. If “yes”, then the client partition proceeds with normal device driver load and configuration 715, which ends processing 720. If the client partition does not have its own device driver for a discovered I/O device, then processing determines whether a partition adjunct image is available for this I/O device 725. If “no”, then the device cannot be configured 730, and processing ends 720.

Assuming that a partition adjunct image is available, then the client partition reserves or donates a portion of its virtual address space and memory space for the partition adjunct image 735, and invokes the hypervisor to instantiate the partition adjunct with the donated virtual address space and memory space, as well as the I/O device 740. The hypervisor creates the new partition adjunct and assigns the donated resources to the partition adjunct 745. The hypervisor then sets up appropriate hardware protection and enforcement of the donated virtual address space and memory space, so that the client partition can no longer access the donated space 750. A new dispatchable state for the partition adjunct is created by the hypervisor 755. As noted above, this new dispatchable state can advantageously be dispatched on a single hardware thread of an SMT data processing system,

while other hardware threads continue to run the invoking (or requesting) client partition. The hypervisor loads the partition adjunct image into the donated adjunct memory 760, and initializes the partition adjunct execution environment 765. The partition adjunct is then dispatched to begin initialization execution at the partition adjunct image entry point 770. The hypervisor sets up partition adjunct service interface calls for the client partition, which include, for example, a virtual input/output request queuing interface for the subject virtualized device class 775. The hypervisor returns to the client partition with the status of the partition adjunct instantiation, and the new virtual device assigned to the client partition 780, and the client partition proceeds with normal virtual device driver load and configuration 785.

FIG. 8 is a flowchart of one embodiment of a process for instantiating a global partition adjunct, such as depicted in FIG. 4B. In this example, the partition adjunct instantiation begins from a logical partition which functions as a virtual input output server (VIOS). The adjunct instantiation setup protocol 800 begins with the VIOS reserving or donating part of its memory space and virtual address space for the global partition adjunct 805. The VIOS invokes the hypervisor to instantiate the global partition adjunct with the donated memory and virtual address space, as well as the I/O device to which accelerated, virtualized access is to be provided 810. The hypervisor creates the new global partition adjunct, again hidden from the administrator, and assigns to the partition adjunct the donated memory space and the I/O device 815. The hypervisor sets up the hardware protection and enforcement of the donated virtual address space and memory space, so that the VIOS can no longer access the donated space 820. The hypervisor creates a new dispatchable state for the partition adjunct 825, which can be dispatched on an individual SMT thread, while one or more other SMT threads of a multithreaded hardware processor core continue to run the VIOS partition invoking the partition adjunct 825. The hypervisor loads the global partition adjunct program image into the donated adjunct memory 830, and initializes the adjunct execution environment 835. The hypervisor then dispatches the global partition adjunct to begin initialization execution at the adjunct image entry point 840, and sets up the global partition adjunct service interface call(s) and virtual device instance(s) available for assignment or attachment from client partitions 845, which completes the processing protocol 850.

FIG. 9 depicts one example of a process for attaching to a global partition adjunct, such as a VIOS partition adjunct created as described above in connection with FIG. 8. This client partition attachment 900 begins with the client partition performing virtual I/O device discovery 905. Processing determines whether the virtual I/O device at issue is capable of global partition adjunct acceleration 910, that is, whether the adjunct service globally exists. If “no”, then the client partition proceeds with normal virtual device configuration 915, for example, using a VIOS, which ends the processing 920. If a virtual I/O device adjunct image is available, then the client partition reserves or donates virtual address space for the adjunct 925, and invokes the hypervisor to attach the global partition adjunct image with the donated virtual address space 930. The hypervisor sets up the hardware protection and enforcement of the donated virtual address space for the client partition 935, and the global partition adjunct is mapped into the client partition page table utilizing the donated virtual address space. The hypervisor creates a virtual I/O mapping for this client partition to the global partition adjunct. Thereafter, processing returns to the client partition to continue with normal virtual device configuration 940, which completes processing 920.

11

FIG. 10 illustrates one embodiment of protocol for processing a service request by a partition adjunct submitted from a client partition. Partition adjunct servicing **1000** begins with inquiring whether the client partition needs to invoke the partition adjunct service **1005**. If “no”, then processing ends **1010**. Otherwise, the client partition invokes a hypervisor call to the partition adjunct service interface for the desired adjunct service call **1015**. Processing determines whether the partition adjunct service is available **1020**. If “no”, then the service request is failed **1025**, and processing ends **1010**.

Assuming that the partition adjunct service is available, then the hypervisor receives the partition adjunct service request **1030**, and determines whether to queue the service request or invoke the service request currently **1035**. If the service request is to be queued, then the hypervisor queues the service request for the partition adjunct **1040**, and returns to inquire whether the client partition needs to invoke a partition adjunct service **1005**. If the service request is to be currently invoked, then the hypervisor performs a “light” context switch of the current state machine, and gives control of the processor to the partition adjunct **1050**. This light context switch is explained in detail below with reference to FIGS. 12-22. The partition adjunct performs or initiates the requested service, and then returns to the hypervisor **1055**. The hypervisor performs a similar light context switch and gives control of the processor back to the client partition initiating the partition adjunct service **1060**, which completes the processing **1010**.

FIG. 11 depicts a flowchart of one embodiment of protocol for asynchronously handling service request completion from a partition adjunct. Partition adjunct request completion processing **1100** begins with an inquiry confirming completion of the partition adjunct service request **1105**. If request processing is not complete, then processing waits until the partition adjunct service request has been completed, at which point the completion of the service request is queued to the hypervisor **1110**. Processing then determines whether the client partitioned is polling or expecting an interrupt the regarding the completion of the service request **1115**. If polling is employed, the client partition polls the hypervisor for completion of the service request **1120**, otherwise the hypervisor posts an interrupt for the client partition **1125**. The hypervisor then performs the light context switch (described below) and invokes the client partition interrupt completion handler **1130**. The client partition processes completion of the service request **1135**, which completes the processing protocol **1140**.

FIGS. 12-15 illustrate one example of a full logical partition to logical partition context switch within a data processing system. When context switching between two separate full partitions, each logical partition has its own distinct virtual address space assigned to it. When the hypervisor performs a full partition switch on an individual processor, the hypervisor must save and re-store processor context machine state (general purpose registers, floating point registers, special purpose registers, instruction address registers, etc.), as well as invalidate and reinitialize the memory management unit (MMU) hardware. This includes invalidating any address translation caches (e.g., translation lookaside buffer), since a full logical partition context switch also means switching to a different virtual address space, and changing the hardware page table pointers in the processor to point to the new logical partition’s in-memory page table. The MMU impacts of a full logical partition context switch are the largest cost in performing a logical partition context switch, as the new logical partition must also re-establish all of its address translations

12

upon reference each time it is dispatched. This is contrasted with the adjunct partition concept disclosed herein (since the adjunct and its client partition share a virtual address space), wherein the MMU hardware does not need to be invalidated or modified, only the subset of processor context/registers that the partition adjunct is allowed to use need to be saved and restored as part of the context switch of the state machine to the partition adjunct.

As illustrated in FIG. 12, a first logical partition (LPAR 1) **1200** includes a virtual address to real address page table **1205**, which comprises the range of virtual addresses assigned to LPAR 1. Similarly, a second logical partition (LPAR 2) **1210** includes a virtual address to real address page table **1215**, which comprises the range of virtual addresses assigned to LPAR 2. The virtual address space range employed by LPAR 1 is separate from the virtual address space employed by LPAR 2, and one logical partition cannot access the virtual address space of the other logical partition.

FIG. 13 illustrates one example of a processor machine state including memory management and address translation hardware of a processor core machine **1300**. The illustrated hardware includes general purpose registers **1310**, floating point registers **1320**, vector registers **1330**, special purpose registers **1340**, a translation lookaside buffer **1350**, and a segment lookaside buffer **1360**. The translation lookaside buffer is a cache of virtual address to real address translations, while the segment lookaside buffer contains effective to virtual address translations.

In FIG. 14, LPAR 1 is assumed to be dispatched, i.e., in an active or running state in the processor state machine. As illustrated, this LPAR 1 running machine state comprises state data in general purpose registers **1310**, floating point registers **1320**, vector registers **1330**, special purpose registers **1340**, translation lookaside buffer **1350**, and segment lookaside buffer **1360**. A special purpose register SDR1 **1400** points to, and therefore, empowers, the virtual address to real address page table of the logical partition in running state, that is, LPAR 1 **1200** in this example. LPAR 2 **1210** is in an inactive state.

FIG. 15 illustrates status of the processor core machine state after LPAR 1 has been switched out and LPAR 2 switched in. As illustrated, the machine state data of LPAR 2 **1210** is restored within general purpose registers **1310**, floating point registers **1320**, vector registers **1330**, and special purpose registers **1340**. The special purpose register SDR1 **1400** in this case is switched to point to the virtual to real address table of LPAR 2 **1210**, with LPAR 1 **1200** being inactive. As a result, the segment lookaside buffer **1360** is restored with LPAR 2 state data, and the translation lookaside buffer **1350** is invalidated. This buffer will fill as LPAR 2 accesses memory addresses.

In contrast, FIGS. 16-18 illustrate context switching of the current machine state from a logical partition to a partition adjunct, in accordance with an aspect of the present invention. FIG. 16 depicts one embodiment of a data processing system comprising a logical partition **1600** (LPAR 1), which includes a virtual address to real address page table **1620**. The data processing system further includes a partition adjunct **1610**. In this example, three page table entries **1621** are set aside to map the partition adjunct into the virtual address space of the logical partition. This example assumes that logical partition **1600** is a client partition which initiates the partition adjunct **1610**. Thus, as illustrated in FIG. 16, the initiating logical partition has donated virtual address space to the partition adjunct, and the partition adjunct virtual address space resides in the same virtual address range as the initiating logical partition. This is represented by the page table entries being

13

set aside for the partition adjunct being within the virtual address to real address page table **1620** of LPAR **1**. As noted above, from the hypervisor's standpoint, in terms of scheduling and isolation, the partition adjunct is a fully separate partition from the initiating logical partition (LPAR **1**). Notwithstanding that, the partition adjunct and the logical partition share a virtual address space range as represented.

FIG. **17** illustrates an example of the processor core machine state with LPAR **1** dispatched in an active, running state. As illustrated, LPAR **1**'s machine state data acquires general purpose registers **1710**, floating point registers **1720**, vector registers **1730**, and special purpose registers **1740**. In addition, the partition's machine state substantially fills the translation lookaside buffer **1750**, and fills the segment lookaside buffer **1760**. Special purpose register SDR1 **1700** points to the virtual address to real address page table **1620** of LPAR **1** **1600**. In this example, partition adjunct **1610** is in the inactive state. As illustrated, translation lookaside buffer **1750** may still contain one or more entries **1751** for the partition adjunct from a prior time that the adjunct was run. These entries **1751** do not need to be invalidated, since logical partition **1600** is protected from accessing those virtual addresses via hardware protection keys used to isolate sub-portions of the single logical partition address space. For example, page protection mechanisms, such as using a protection key, controlled by a special purpose register, labeled access mask register (AMR) **1770** may be employed. The AMR register **1770** may hold keys that are actively turned on at any given moment. For a page table entry to be referenced, the associated keys must be enabled in the AMR register. When the hypervisor is running LPAR **1**, the hypervisor sets the protection key so that LPAR **1** cannot get access to the page table entries **1621**, or the entries **1751** of translation lookaside buffer **1750**. Further details of this protection and isolation scheme are described in the above-incorporated, co-filed application entitled "Hypervisor-Enforced Isolation of Entries Within a Single Logical Partition's Virtual Address Space".

FIG. **18** illustrates the processor core machine state with partition adjunct **1610** dispatched active. In this example, the adjunct service provided by partition adjunct **1610** employs general purpose registers **1710**, but not floating point registers **1720** and vector registers **1730**, which retain the state information for LPAR **1** **1600**, presently in the inactive state. Further, only selected special purpose registers **1740** are employed by the partition adjunct, with the remaining special purpose registers being invalidated or containing state information for LPAR **1** **1600**. The special purpose register SDR1 **1700** (in this example) continues to point to the virtual address to real address page table **1620** of LPAR **1** since partition adjunct **1610** is mapped into the donated virtual address spaces **1621** thereof. Various entries **1751** in translation lookaside buffer **1750** are being employed with the partition adjunct dispatched, and the remaining entries retain state information of LPAR **1**. A single entry **1761** in segment lookaside buffer **1760** is installed to cover the adjunct's portion of the address space, with the remaining entries retaining the state information for LPAR **1**, which is protected from access by LPAR **1**. Because of the protection keys employed, for example, by special purpose register AMR **1770**, the partition adjunct only has access to the noted registers and table entries. From a hypervisor standpoint, dispatching of the partition adjunct is separate from dispatching of the LPAR **1**, however, the processor core machine state information is held for LPAR **1** in various registers and tables not employed by the partition adjunct. The result is that the context switch from LPAR **1** to the partition adjunct is a "light" context

14

switch of the state machine wherein only selected registers and table information is updated with the partition adjunct state information. Similarly, when switching from the partition adjunct back to LPAR **1**, only the register information employed by the partition adjunct needs to be updated for translation back to LPAR **1** active state.

FIGS. **19-22** illustrate one example of context switching the current state machine between a logical partition and a partition adjunct, such as described herein, in a data processing system employing simultaneous hardware multithreading. The example depicted is a two hardware-threaded core. However, those skilled in the art will understand that the concepts described are readily applicable to more than a two hardware-threaded core. By way of example, the POWER5™ and POWER6™ Power Instruction Set Architecture Processors, offered by International Business Machines Corporation, of Armonk, N.Y., employ two hardware-threaded cores.

As illustrated in FIG. **19**, the multithreaded core includes a first hardware thread **1900** and a second hardware thread **1910** which share a translation lookaside buffer **1920**. Each hardware thread includes, for example, the general purpose registers, floating point registers, and special purpose registers, as well as the segment lookaside buffer described above in connection with FIGS. **13-18**. In FIG. **19**, the multithreaded hardware core is illustrated without a logical partition being active. In FIG. **20**, a first logical partition (LPAR **1**) is assumed to be active, with the LPAR **1** state information occupying the hardware-threaded cores **1900**, **1910**, as well as the table lookaside buffer **1920**.

In FIG. **21**, a context switch of the state machine has occurred between the first logical partition (LPAR **1**) and a second logical partition (LPAR **2**). As illustrated, this is a full context switch, wherein LPAR **2**'s machine state occupies first hardware-threaded core **1900**, second hardware-threaded core **1910**, and table lookaside buffer **1920** of this two hardware-threaded core example.

FIG. **22** illustrates an example of context switching from LPAR **1** to a partition adjunct initiated by LPAR **1**. As illustrated, only a single thread **1910** is updated with partition adjunct's state information, i.e., undergoes a light context switch. Specifically, the general purpose registers and various special purpose registers are employed in this example by the partition adjunct service, with the remaining registers and the segment lookaside buffer retaining machine state information of the logical partition (LPAR **1**). Further, the second thread, that is, hardware thread **1900**, retains and remains running in the initiating logical partition (i.e., LPAR **1**). In this example, the machine state for LPAR **1** is also retained by the table lookaside buffer, with only selected entries **1921** being employed by the partition adjunct. Each hardware thread has its own access mask register that contains a set of keys which control what hardware the partition adjunct (when active on that thread) has access to. Thus, LPAR **1** remains active on one thread, while the partition adjunct is active on a different thread of the multithreaded data processing system.

In a further aspect, client partition scheduling and prioritization control over a service partition's work is described below with reference to FIG. **23**. By way of example, the service partition is a partition adjunct characterized as described above. In an alternate implementation, the service partition may comprise any service logical partition, such as a virtual input/output server (VIOS) partition, which is characterized as lacking its own assigned processor resources, as explained further below.

Conventionally, a data processing system employing a virtual input/output service partition requires that the hardware system contain at least two logical partitions (LPARs); that is,

a client partition and a server partition. This in turn requires that the hypervisor treat at least one processor (or CPU) as shared, and provide dispatched quanta to the two logical partitions at regular intervals. This effectively prevents the client partition's operating system from employing any internal scheduling or prioritization mechanisms to control processor usage associated with the input/output service. The control essentially resides in the hypervisor, which is forced to share the processor(s) between the two logical partitions' virtual processors. In such an implementation, each logical partition is a full, distinct operating environment which requires its own set of system resources, including processor and memory. Each partition gets dispatched or scheduled onto physical resources by the hypervisor, with normal dispatching at regular intervals. Thus, the client partition can post work for the server partition to perform, however, the hypervisor schedules when the work is to be performed by the two partitions, based on pre-assigned resource usage rules. There can be no dynamic, intelligent control of the processing of an input/output service request by the server partition in such a system.

Disclosed hereinbelow is a processing environment and a set of mechanisms and procedures that permit a client partition to determine when a server partition requires processor dispatch, and enables the client partition to provide processor quantum (e.g., CPU cycles) to the service partition as the client partition desires, according to its own internal scheduling and prioritization mechanisms. As one specific example, the server partition is a partition adjunct created and characterized as described above. Alternatively, the service partition could comprise any logical partition characterized as lacking its own assigned processor resources.

Assuming that the server partition is a partition adjunct, one property of which is that the partition adjunct does not require regular dispatches in the same manner as a general purpose operating system, then the hypervisor does not explicitly share the processor between the client partition and the service partition. Furthermore, explicit controls (such as those described below) can be employed to provide the client partition's operating system with control over how processor time for input/output services is employed.

If virtual devices are used to provide access to input/output services on a virtualization-enabled platform, then the drivers needed to drive the physical input/output adapter will reside on the service partition, and not within the client partition. Conventionally, as noted, the client of these services has no ability to control how much processor time is granted to computation in support of input/output services. Conversely, if the operating system of the client partition employs a native device driver, then it can apply various prioritization and scheduling schemes to achieve processor time constraints on input/output servicing related computation. With a device driver located in a service partition, however, control of scheduling resides in the service partition's operating system and the hypervisor. The solution presented herein is to enable the client operating system to have control over the service partition's access to computational resources and computational time, for example, by implementing the service partition as a partition adjunct. Multiple services may be offloaded from a client partition by providing multiple partition adjuncts, each implementing a respective service for the client partition.

As noted above, a partition adjunct, unlike a general purpose operating system, does not require regular processor dispatch; that is, a partition adjunct can be dispatched only

when there is work for it to do, or if an interrupt or timeout requiring servicing has been noted, for example, to indicate an error condition.

In accordance with an aspect of the present invention, a hypervisor permits a client partition to perform directed yield or handoff operations to a service partition (or a virtual service provider) characterized as described herein, and for the hypervisor to communicate to the client partition that such a server partition is runnable with a virtual interrupt that is raised to indicate this condition. Client partitions may thus use the directed-yield facilities to adjust the relative priority of their service partitions' processes, without affecting the scheduling policies imposed on the client partition(s) by the system administrator. For example, a client partition cannot use this mechanism to increase the total number of processor cycles assigned to it, or other entities running on its behalf.

In one implementation, the hypervisor maintains a constant scheduling policy assigning processor time slices to a logical partition (LPAR), and the logical partition may not manipulate the hypervisor so as to obtain more processor time than has been allocated to it. However, the logical partition may use the mechanisms described herein to direct extra processor time (or quanta) to one of its service partitions (as per its internal scheduling and prioritization policies). With this protocol, it is possible that the service partition may receive an extra allocation of processor time due to donations of processor time to it, however, such a service partition is exclusive to a particular client partition (and thus, its processor time is considered processor time of the client partition), or if it is shared, then the service partition is a shared, trusted platform component that can be trusted with the processor time provided to it.

FIG. 23 depicts one embodiment of a process for client partition control of prioritizing and scheduling service requests for execution by a service partition of a data processing system. In this embodiment, both the client partition (i.e., client LPAR) and the service partition (e.g., partition adjunct) run above and interact through the hypervisor of the data processing system.

Processing begins with a client operating system **2300** of a client LPAR generating a service request which is initially queued for processing by the partition adjunct (or more generally, the service partition) **2305**. Processing determines whether there are more service requests **2310**, and if "yes", continues to queue or batch the service requests **2305**. If there are currently no additional service requests, then processing determines whether the partition adjunct should now be scheduled (or called) for execution of one or more of the service requests **2315**. This scheduling inquiry evaluates priority of the service requests, in comparison with priority of other work the client operating system is performing or is to perform. Assuming that the partition adjunct is not to be currently scheduled or called, then the client operating system performs the other higher priority work **2320**, and subsequently returns to again determine whether the partition adjunct can be scheduled **2315**.

Once the client operating system determines that the partition adjunct can be scheduled, then a call is made to the hypervisor to yield the client LPAR's processor resource to the partition adjunct **2325**. Responsive to this call, the hypervisor saves the client LPAR state (or thread state in a multi-threaded data processing system implementation), and dispatches the partition adjunct on the calling processor resource (e.g., the calling processor thread) **2330**. The partition adjunct **2335** initiates processing by determining whether there is any pending service request work **2340**. If "yes", then the partition adjunct processes the pending service request(s) **2345**.

Once all service request work has been processed, the partition adjunct yields the processor resource back to the client LPAR. Specifically, the hypervisor saves the partition adjunct's state, and resumes the client LPAR state where previously saved **2350**, which returns processing back to the client LPAR **2355** and completes the directed yield of resources to the partition adjunct for processing of the service request (e.g., an input/output service request) **2360**.

As noted above, along with yielding the client partition's processor resource to the partition adjunct for servicing of the service request, the partition adjunct receives a processor quantum (i.e., time interval) in which to employ the processor resource. The hypervisor includes processing for monitoring whether the processor quantum has expired **2365** (for example, a hypervisor timer would be set to expire at the end of the allotted quantum and preempt the partition adjunct). If "yes", then a partition busy interrupt signal (i.e., "PA BUSY") **2375** is posted by the hypervisor for the client LPAR. In this case, the hypervisor saves the partition adjunct thread state, and dispatches the client LPAR at the busy interrupt vector **2380**. The client LPAR begins processing at the partition busy interrupt signal **2385**, and determines whether to currently schedule or call the partition adjunct for further processing of the service request **2315**, or whether to perform other higher priority work at the client LPAR **2320** before scheduling (or calling) the partition adjunct for further work.

In addition, the hypervisor monitors for an asynchronous interrupt or timeout parameter requiring servicing **2370**, which may be (for example) a course-grain timer to indicate that an error condition may have occurred. If such an interrupt or timer expiration preempts the partition adjunct, then the hypervisor posts a partition busy interrupt signal for the client LPAR **2375**, and processing proceeds as described above.

Advantageously, those skilled in the art will note from the above description that a service partition (such as a partition adjunct as described herein) operates on processor (i.e., CPU) resources of the request originator, that is, the client partition. The partition adjunct does not require regular dispatching by the hypervisor, and is dispatched only on demand, as needed, for service request processing or completion of service request processing. A client partition or client operating system utilizing one or more partition adjuncts can decide when and if it wants to donate its processor resource(s) and processor quantum to the partition adjunct servicing one or more requests, such as input/output requests of the client. This allows for efficient prioritization and scheduling of processor resources by the client partition itself.

Further, the client partition may receive a virtual busy interrupt signal from its partition adjunct, with the interrupt signal being a unique vector to a particular virtual service partition, e.g., to a particular virtual input/output service. When a partition adjunct requires additional processor time, either due to an interrupt or because it has exceeded its allotted dispatch quantum without entering a quiescent state, the hypervisor is signaled to raise the partition busy interrupt signal to the client partition. The client partition indicates to the hypervisor that the partition adjunct is to run at a time of the client partition's choosing, and at that time, the client partition performs a directed yield operation to the partition adjunct to donate, for example, the rest of a further scheduled processor quantum of the client partition.

One or more aspects of the present invention can be included in an article of manufacture (e.g., one or more computer program products) having, for instance, computer usable media. The media has therein, for instance, computer readable program code means or logic (e.g., instructions, code, commands, etc.) to provide and facilitate the capabilities

of the present invention. The article of manufacture can be included as a part of a computer system or sold separately.

One example of an article of manufacture or a computer program product incorporating one or more aspects of the present invention is described with reference to FIG. **24**. A computer program product **2400** includes, for instance, one or more computer usable media **2410** to store computer readable program code means or logic **2420** thereon to provide and facilitate one or more aspects of the present invention. The medium can be an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system (or apparatus or device) or a propagation medium. Examples of a computer readable medium include a semiconductor or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk and an optical disk. Examples of optical disks include compact disk-read only memory (CD-ROM), compact disk-read/write (CD-R/W) and DVD.

A sequence of program instructions or a logical assembly of one or more interrelated modules defined by one or more computer readable program code means or logic direct the performance of one or more aspects of the present invention.

Although various embodiments are described above, these are only examples.

Moreover, an environment may include an emulator (e.g., software or other emulation mechanisms), in which a particular architecture or subset thereof is emulated. In such an environment, one or more emulation functions of the emulator can implement one or more aspects of the present invention, even though a computer executing the emulator may have a different architecture than the capabilities being emulated. As one example, in emulation mode, the specific instruction or operation being emulated is decoded, and an appropriate emulation function is built to implement the individual instruction or operation.

In an emulation environment, a host computer includes, for instance, a memory to store instructions and data; an instruction fetch unit to fetch instructions from memory and to optionally, provide local buffering for the fetched instruction; an instruction decode unit to receive the instruction fetch unit and to determine the type of instructions that have been fetched; and an instruction execution unit to execute the instructions. Execution may include loading data into a register for memory; storing data back to memory from a register; or performing some type of arithmetic or logical operation, as determined by the decode unit. In one example, each unit is implemented in software. For instance, the operations being performed by the units are implemented as one or more subroutines within emulator software.

Further, a data processing system suitable for storing and/or executing program code is usable that includes at least one processor coupled directly or indirectly to memory elements through a system bus. The memory elements include, for instance, local memory employed during actual execution of the program code, bulk storage, and cache memory which provide temporary storage of at least some program code in order to reduce the number of times code must be retrieved from bulk storage during execution.

Input/Output or I/O devices (including, but not limited to, keyboards, displays, pointing devices, DASD, tape, CDs, DVDs, thumb drives and other memory media, etc.) can be coupled to the system either directly or through intervening I/O controllers. Network adapters may also be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote printers or storage devices through intervening private or public net-

19

works. Modems, cable modems, and Ethernet cards are just a few of the available types of network adapters.

The capabilities of one or more aspects of the present invention can be implemented in software, firmware, hardware, or some combination thereof. At least one program storage device readable by a machine embodying at least one program of instructions executable by the machine to perform the capabilities of the present invention can be provided.

The flow diagrams depicted herein are just examples. There may be many variations to these diagrams or the steps (or operations) described therein without departing from the spirit of the invention. For instance, the steps may be performed in a differing order, or steps may be added, deleted, or modified. All of these variations are considered a part of the claimed invention.

Although embodiments have been depicted and described in detail herein, it will be apparent to those skilled in the relevant art that various modifications, additions, substitutions and the like can be made without departing from the spirit of the invention and these are therefore considered to be within the scope of the invention as defined in the following claims.

What is claimed is:

1. A method in a data processing system for processing a service request, the method comprising:

obtaining by a service partition of the data processing system a service request from a logical partition of the data processing system, the logical partition having been created by and executing above a hypervisor of the data processing system, and the hypervisor having assigned a processor quantum to the logical partition, and the service partition also having been created by and executing above the hypervisor, wherein the hypervisor enforces security and isolation between the service partition and the logical partition, such that the service partition lacks access to memory management unit configuration and floating-point facilities accessible to the logical partition, and the service partition can access functionality needed to run the service request; and

processing the service request by the service partition utilizing the processor quantum assigned to the logical partition by the hypervisor and donated by the logical partition to the service partition.

2. The method of claim 1, further comprising controlling by the logical partition scheduling of the service partition, the controlling comprising queuing the service request at the logical partition until the logical partition decides to schedule execution of the service request by the service partition.

3. The method of claim 2, wherein the controlling further comprises queuing a plurality of service requests at the logical partition before the logical partition decides to schedule execution of one or more service requests of the plurality of queued service requests by the service partition.

4. The method of claim 2, further comprising delaying by the logical partition scheduling of execution of the service request by the service partition pending execution of higher priority work by the logical partition.

5. The method of claim 2, further comprising, responsive to the logical partition deciding to schedule execution of the service request, calling by the logical partition the hypervisor to yield a processor resource assigned to the logical partition and donate the processor quantum to the service partition for processing of the service request, and dispatching by the hypervisor the service partition on the processor resource for execution of the service request utilizing the donated processor quantum.

20

6. The method of claim 5, further comprising calling by the service partition the hypervisor to yield the processor resource back to the logical partition upon completion of processing of the service request, and dispatching by the hypervisor the logical partition on the processor resource for resuming execution of the logical partition utilizing the processor resource.

7. The method of claim 1, wherein the service partition processes the service request utilizing a virtual address space assigned by the hypervisor to the logical partition and donated to the service partition by the logical partition.

8. The method of claim 1, further comprising monitoring by the service partition for expiration of the processor quantum prior to completion of processing of the service request, and when detected, automatically posting a partition busy interrupt signal by the hypervisor to the logical partition, saving service partition state information by the hypervisor, and dispatching the logical partition to process the partition busy interrupt signal, wherein processing the partition busy interrupt signal by the logical partition comprises determining whether to currently schedule the service request for further execution by the service partition.

9. The method of claim 8, wherein processing the partition busy interrupt signal by the logical partition further comprises automatically determining whether processing of the service request by the service partition is of higher priority than other work to be performed by the logical partition.

10. The method of claim 1, further comprising monitoring by the hypervisor for receipt of an interrupt or timeout parameter requiring servicing by the service partition on behalf of the logical partition, and when detected, automatically posting a partition busy interrupt signal by the hypervisor for the logical partition, saving service partition state information by the hypervisor, and dispatching the logical partition to process the partition busy interrupt signal, wherein processing the partition busy interrupt signal comprises determining by the logical partition whether to currently schedule the service partition for processing of the interrupt or timeout parameter requiring servicing, or whether to delay processing of the interrupt or timeout parameter requiring servicing, pending execution of higher priority work by the logical partition.

11. A data processing system comprising:

at least one processing environment comprising a logical partition and a service partition to process a service request for the logical partition, the logical partition having been created by and executing above a hypervisor of the data processing system, the hypervisor having assigned a processor quantum to the logical partition, and the service partition also having been created by and executing above the hypervisor;

wherein the hypervisor enforces security and isolation between the service partition and the logical partition, such that the service partition lacks access to memory management unit configuration and floating-point facilities accessible to the logical partition, and the service partition can access functionality needed to run the service request; and

wherein the service request is processed by the service partition utilizing the processor quantum assigned to the logical partition by the hypervisor and donated by the logical partition to the service partition with a call to the service partition to process the service request.

12. The data processing system of claim 11, wherein the service partition processes the service request utilizing a virtual address space assigned by the hypervisor to the logical partition and donated to the service partition by the logical partition.

21

13. The data processing system of claim 11, wherein the logical partition controls scheduling of the service partition, and wherein control of scheduling by the logical partition includes queuing the service request at the logical partition until the logical partition decides to schedule execution of the service request by the service partition.

14. The data processing system of claim 13, wherein the logical partition delays scheduling of execution of the service request by the service partition pending execution of higher priority work by the logical partition.

15. The data processing system of claim 13, wherein the hypervisor monitors for expiration of the processor quantum prior to completion of processing of the service request by the service partition, and when detected, automatically posts a partition busy interrupt signal for the logical partition, and the hypervisor saves service partition state information, and dispatches the logical partition to process the partition busy interrupt signal, wherein processing the partition busy interrupt signal comprises determining whether to currently schedule the service partition for further execution of the service request.

16. The data processing system of claim 11, wherein the data processing system comprises a hardware multithreaded data processing system, and wherein the hypervisor dispatches the service partition on a single thread of the hardware multithreaded data processing system, while at least one other thread of the hardware multithreaded data processing system continues to run in the logical partition issuing the service request.

17. The data processing system of claim 11, wherein the hypervisor monitors for receipt of an interrupt or timeout parameter requiring servicing by the service partition on behalf of the logical partition, and when detected, automatically posts a partition busy interrupt signal for the logical partition, and saves service partition state information and dispatches the logical partition to process the partition busy interrupt signal, wherein the logical partition processes the partition busy interrupt signal by determining whether to currently schedule the service partition for processing of the interrupt or timeout parameter requiring servicing, or whether to delay processing of the interrupt or timeout parameter requiring servicing, pending execution of higher priority work by the logical partition.

22

18. An article of manufacture comprising:

at least one non-transitory computer-usable medium having computer-readable program code logic defining a method of processing a service request within a data processing system, the computer-readable program code logic when executing on a processor performing:

obtaining by a service partition of the data processing system a service request from a logical partition of the data processing system, the logical partition having been created by and executing above a hypervisor of the data processing system, and the hypervisor having assigned a processor quantum to the logical partition, and the service partition also having been created by and executing above the hypervisor, wherein the hypervisor enforces security and isolation between the service partition and the logical partition, such that the service partition lacks access to memory management unit configuration and floating-point facilities accessible to the logical partition, and the service partition can access functionality needed to run the service request; and

processing the service request by the service partition utilizing the processor quantum assigned to the logical partition by the hypervisor and donated by the logical partition to the service partition.

19. The article of manufacture of claim 18, wherein the service partition processes the service request utilizing a virtual address space assigned by the hypervisor to the logical partition and donated to the service partition by the logical partition.

20. The article of manufacture of claim 18, wherein the performing further comprises monitoring by the service partition for expiration of the processor quantum prior to completion of processing of the service request by the service partition, and when detected, automatically posting a partition busy interrupt signal by the hypervisor for the logical partition, saving service partition state information by the hypervisor, and dispatching the logical partition to process the partition busy interrupt signal, wherein processing the partition busy interrupt signal by the logical partition comprises determining whether to currently schedule the service request for further execution by the service partition.

* * * * *